The Dissertation Committee for Rong Wang Certifies that this is the approved version of the following dissertation:

Proteomic Analysis of Mycobacteria and Mammalian Cells

Committee:

Edward M. Marcotte, Supervisor

Vishwanath R. Iyer

David W. Hoffman

Barrie G. Kitto

Jennifer S. Brodbelt

Proteomic Analysis of Mycobacteria and Mammalian Cells

by

Rong Wang, B.S.; M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May, 2006

Dedication

To my dear parents.

Acknowledgements

I am very grateful for my doctorate advisor, Dr. Edward M. Marcotte. His endorsement in science impressed me a lot. It would not have been possible to get it finished without the direction and support from him. I would like to acknowledge all committee members who gave me thoughtful suggestions on my projects. I thank all colleagues in the lab for the friendly environment they created and scientific suggestions they provided. Particular appreciation goes to John Prince who helped me a lot in Mass Spectrometry technology understanding and experiments. Dr. Jian Gu gave me scientific discussion on statistical methods. Kris Mcgary and Mark Carlson proposed corrections and proofread the manuscript. David Parr helped me with Mass Spec data backup. Finally, I would like to thank my husband and friends for their care and support.

Proteomic Analysis of Mycobacteria and Mammalian Cells

Publication No._____

Rong Wang, Ph.D.

The University of Texas at Austin, 2006

Supervisor: Edward M. Marcotte

Tuberculosis is a serious threat that claims 2 million lives annually. *Mycobacterium tuberculosis* is the causative agent of tuberculosis. The fast-growing bacteria *Mycobacterium smegmatis* is a model mycobacterial system, a non-pathogenic soil bacterium that nonetheless shares many features with the pathogenic *M. tuberculosis*. Multidimensional chromatography coupled with shotgun style tandem mass spectrometry was used to detect and identify 2,550 distinct proteins from *M. smegmatis* with an estimated 5% false positive identification rate, many predicted genes were annotated using experimental results and protein expression levels were estimated from the shotgun proteomic data.

First, in 25 exponential and stationary phase experiments, we observed numerous proteins involved in energy production, protein translation, and lipid biosynthesis. Protein expression levels were estimated from the number of observations of each protein, allowing measurement of differential expression of complete operons, and the

comparison of the stationary and exponential phase proteomes. Expression levels are correlated with proteins' codon biases and mRNA expression levels.

Secondly, we measured changes in the proteome of *Mycobacterium smegmatis* in response to three anti-tuberculosis drugs isoniazid (INH), ethambutol (EMB) and 5-chloro-pyrazinamide (5-Cl-PZA). Protein expression levels were calculated from the number of identified peptides for each protein. Translation, energy production, and protein export are all down-regulated in the three drug treatments. By contrast, systems related to drugs' targets, including lipid, amino acid, nucleotide metabolism and transport, show specific protein expression changes associated with each drug treatment. We use these changes to infer likely targets for PZA.

Thirdly, computational methods were used to predict protein-protein interactions and protein functions in a metabolic pathway in *M. tuberculosis*. Protein functional links were built and specific functions were characterized for the pathway and its parallel pathways in *M. tuberculosis* and other organisms.

Finally, multidimensional chromatography coupled with shotgun style tandem mass spectrometry has been applied in the analysis of nuclear proteins from mammalian cells. Nuclear proteins were identified from mouse T lymphoma cells. Nuclear matrix-associated proteins were identified from human preliminary T cells during the transition from quiescent state to proliferating state. These proteins are involved in the function of DNA replication, RNA transcription, splicing, etc.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Tuberculosis (TB) is one of the most prevalent infectious diseases in the world. Annually, there are eight million new cases and two million fatalities, according to the World Health Organization (http://www.who.int/tb/en). The cause of TB, *Mycobacteria tuberculosis*, is a slow-growing, aerobic gram-positive bacterium (Cole et al. 1998b). Although current first-line anti-TB drug regimens can achieve more than 99% efficacy, lengthy chemotherapy and emergence of drug-resistant strains pose significant problems for effective control (Corbett et al. 2003). Understanding the biological action of the currently used drugs will facilitate the discovery of new drugs that shorten and simplify the treatment of TB.

The sequencing of the *M. tuberculosis* genome has accelerated progress in understanding TB (Cole et al. 1998b). DNA microarray technology has been used (Boshoff et al. 2004; Talaat et al. 2002; Wilson et al. 1999) to explore the function of genes and interactions between genes in *M. tuberculosis* under different growth conditions. As a complement to genomics data, proteomics has been widely used because biological activity is carried out mainly by the dynamic population of proteins (de Hoog and Mann 2004). A major goal of proteomics is quantitative analysis of all the proteins expressed in a cell or an organism under a specific condition (Pandey and Mann 2000). The proteome reflects the interaction and regulation between proteins and the functional status of a cell in response to environmental stimuli (Ong and Mann 2005).

Proteomics is built on technologies that analyze large numbers of proteins —— ideally the entire proteome in the same experiment. Two-dimensional polyacrylamide gel electrophoresis (2DE) is the original high resolution protein separation method available; mass spectrometry is the most versatile technology to directly measure endogenous proteins. Two-dimensional gel electrophoresis coupled with mass spectrometry (2DE-MS) was a popular analytical method used for profiling complete proteomes. Bypassing potential limitations of gel electrophoresis and protein insolubility, shotgun proteomics is a gel-free approach based on multidimensional liquid chromatography coupled to tandem mass spectrometry (LC/LC/MS/MS) (Washburn et al. 2001). In this method, protein mixtures are digested with proteases and the resulting peptides are separated by multi-dimensional liquid chromatography, then the separated peptides are fragmented in the mass spectrometer and the MS/MS spectra are matched to predicted peptide sequences from a database (Link et al. 1999). Protein identification from the complex will be accomplished using SEQUEST (Eng 1994) or Mascot (Perkins et al. 1999) algorithm.

Shotgun proteomics is capable of investigating the systematic response of mycobacteria in a reasonably comprehensive manner (Washburn and Yates 2000). The fast-growing *Mycobacterium smegmatis*, a non-pathogenic mycobacterial model system that shares many features with *M. tuberculosis*, has been used for the development of anti-mycobacterial therapy (Rapaport et al. 1996). The *M. smegmatis* grows faster is because there are two rRNA operons, while only one for *M. tuberculosis* (Ratledge 1999). We applied LC/LC/MS/MS to characterize the expressed proteome of *M. smegmatis*, detect and identify approximately 2,550 distinct proteins, providing experimental annotation for many predicted genes with an estimated 5% false positive identification rate. In Chapter 2, we report the observation of 901 distinct proteins under

25 differing growth conditions, estimate protein expression levels from the number of observations of each protein, and demonstrate that the protein expression levels correlate with codon choice and mRNA expression levels.

Proteomics is not only identification of proteins but characterization of protein expression levels. Quantitative proteomics provides an efficient way to accelerate the discovery and develop novel methods for disease therapy by systematically studying the proteome (Ong and Mann 2005). Changes in protein expression due to stimuli or conditioning are measured in a systematic manner and are used for elucidating mechanisms of cell function and signaling. A number of quantitative methods have been developed to measure protein expression levels from MS/MS data. Using stable isotopes as internal reference standards opened a new era for quantitative proteomics, such as Stable Isotope Labeling by Amino acids in Cell culture (SILAC) (Oda et al. 1999; Ong et al. 2002) and Isotope-Coded Affinity Tag Technology (ICAT) (Gygi et al. 1999). These methods provide a way to obtain relative quantification: each protein's expression change is measured relative to the abundance of isotope labeled reference. In contrast, AQUA (Gerber et al. 2003) provides absolute quantification with the help of known amount of reference peptides spiked into a sample. These methods rely on peptide peak volume measurements for quantification. Absolute protein quantification can also be obtained from counts of the number of MS/MS spectra per protein, which can be related to protein expression level (Ishihama et al. 2005; Liu et al. 2004; Lu et al. 2005a). These emerging techniques for monitoring differential gene expression offer valuable guidance in elucidating regulatory mechanisms of metabolic pathways and thereby pinpointing new drug targets. This idea has been explored in our study of proteomic analysis of mycobacteria in response to anti-TB drugs. In Chapter 3, we treated *Mycobacterium*

smegmatis with anti-TB drugs Isoniazid (INH), ethambutol (EMB), 5-chloro-pyrazinamide (5-Cl-PZA) (Zhang 2005), measured drug-induced alterations in *M. smegmatis* protein profiles, and identified markers of drug action. Protein differential expression levels were estimated from the shotgun proteomic data, and both individual proteins and families of proteins were identified that show altered expression levels in response to drugs. We provide a large-scale analysis of the *M. smegmatis* proteomic response to INH, EMB and 5-Cl-PZA and elucidate the drugs' systematic effects on mycobacterial cells.

Systematic characterization of protein functions and relationships can be obtained not only from the genome-wide experiments, such as shotgun experiments, but also from the information in genome sequences, such as patterns of gene fusion, conservation of gene order, patterns of gene co-inheritance and other sorts of evolutionary information, which can be revealed using computational methods (Eisenberg et al. 2000; Lee et al. 2004; Marcotte et al. 1999b). In Chapter 4, we combined the methods of phylogenetic profiles, Rosetta Stone links, gene neighbor analysis and operon predictions to discover functional links between non-homologous proteins. These techniques are integrated with BLASTP (Altschul et al. 1990) and structure-derived protein functions to predict protein-protein interactions, functions and metabolic pathways. We plotted a metabolic pathway in *M. tuberculosis* and assigned protein functions to the components of the pathway and the parallel pathways in *M. tuberculosis* and other organisms. This study on mycobacteria metabolic pathways sheds light on mechanisms of mycobacterial growth and drug responses, and provides valuable perspectives regarding anti-mycobacterial drug development.

Finally, in Chapter 5, we applied the LC/LC/MS/MS technique to the proteomic analysis of mammalian cells. In mouse T lymphoma cells, nuclear proteins were identified. In human primary T lymphocytes, the matrix-associated proteins involved in the transition from quiescent state to proliferating state were identified. These proteins are involved in the function of DNA replication, RNA transcription, and RNA splicing. These studies provide guidance in understanding how the nucleus functions and the role of nuclear matrix-associated proteins in cell activation.

REFERENCES

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Boshoff, H.I., T.G. Myers, B.R. Copp, M.R. McNeil, M.A. Wilson, and C.E. Barry, 3rd. 2004. The transcriptional responses of Mycobacterium tuberculosis to inhibitors of metabolism: novel insights into drug mechanisms of action. *J Biol Chem* **279:** 40174-40184.

Cole, S.T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry, 3rd, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M.A. Quail, M.A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J.E. Sulston, K. Taylor, S. Whitehead, and B.G. Barrell. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393:** 537-544.

Corbett, E.L., C.J. Watt, N. Walker, D. Maher, B.G. Williams, M.C. Raviglione, and C. Dye. 2003. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med* **163:** 1009-1021.

de Hoog, C.L. and M. Mann. 2004. Proteomics. *Annu Rev Genomics Hum Genet* **5:** 267-293.

Eisenberg, D., E.M. Marcotte, I. Xenarios, and T.O. Yeates. 2000. Protein function in the post-genomic era. *Nature* **405:** 823-826.

Eng, J.K., A.L. McCormack, and J.R.3rd. Yates. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5:** 976-989.

Gerber, S.A., J. Rush, O. Stemman, M.W. Kirschner, and S.P. Gygi. 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **100:** 6940-6945.

Gygi, S.P., B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, and R. Aebersold. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17:** 994-999.

Ishihama, Y., Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, and M. Mann. 2005. Exponentially Modified Protein Abundance Index (emPAI) for Estimation of

Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. *Mol Cell Proteomics* **4:** 1265-1272.

Lee, I., S.V. Date, A.T. Adai, and E.M. Marcotte. 2004. A probabilistic functional network of yeast genes. *Science* **306:** 1555-1558.

Link, A.J., J. Eng, D.M. Schieltz, E. Carmack, G.J. Mize, D.R. Morris, B.M. Garvik, and J.R. Yates, 3rd. 1999. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **17:** 676-682.

Liu, H., R.G. Sadygov, and J.R. Yates, 3rd. 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76:** 4193-4201.

Lu, P., C. Vogel, and E.M.E. Marcotte. 2005. An estimate of relative contributions of transcriptional and translational regulation by absolute protein expression profiling. *submitted*.

Marcotte, E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* **402:** 83-86.

Oda, Y., K. Huang, F.R. Cross, D. Cowburn, and B.T. Chait. 1999. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci U S A* **96:** 6591-6596.

Ong, S.E., B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, and M. Mann. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1:** 376-386.

Ong, S.E. and M. Mann. 2005. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* **1:** 252-262.

Pandey, A. and M. Mann. 2000. Proteomics to study genes and genomes. *Nature* **405:** 837-846.

Perkins, D.N., D.J. Pappin, D.M. Creasy, and J.S. Cottrell. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20:** 3551-3567.

Rapaport, E., A. Levina, V. Metelev, and P.C. Zamecnik. 1996. Antimycobacterial activities of antisense oligodeoxynucleotide phosphorothioates in drug-resistant strains. *Proc Natl Acad Sci U S A* **93:** 709-713.

Ratledge, C.a.D., J. 1999. *Mycobacteria molecular biology and virulence*. Blackwell science.

Talaat, A.M., S.T. Howard, W.t. Hale, R. Lyons, H. Garner, and S.A. Johnston. 2002. Genomic DNA standards for gene expression profiling in Mycobacterium tuberculosis. *Nucleic Acids Res* **30:** e104.

Washburn, M.P., D. Wolters, and J.R. Yates, 3rd. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19:** 242-247.

Washburn, M.P. and J.R. Yates, 3rd. 2000. Analysis of the microbial proteome. *Curr Opin Microbiol* **3:** 292-297.

Wilson, M., J. DeRisi, H.H. Kristensen, P. Imboden, S. Rane, P.O. Brown, and G.K. Schoolnik. 1999. Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization. *Proc Natl Acad Sci U S A* **96:** 12833-12838.

Zhang, Y. 2005. The magic bullets and tuberculosis drug targets. *Annu Rev Pharmacol Toxicol* **45:** 529-564.

Chapter 2. Mass-spectrometry of the *M. smegmatis* proteome: protein expression levels correlate with function, operons, and codon bias

The fast-growing non-pathogenic bacterium *Mycobacterium smegmatis* is particularly useful in studying basic cellular processes of relevance to pathogenic mycobacteria, such as the related species *M. tuberculosis*, the causative agent of tuberculosis. Although the genome sequencing of *M. smegmatis* is nearly complete (http://www.tigr.org) (Brosch *et al.* 2001), much is unknown about the mechanisms controlling growth in mycobacterial species. The large-scale study of the proteins expressed by *M. smegmatis* in different growth states has the potential to generate information about the mechanisms of cell growth, division, and adaptation, as well as inform about mycobacterial proteomes in general.

Until recently, the method of choice for profiling a complete proteome was two-dimensional gel electrophoresis coupled with mass spectrometry (2DE-MS). For example, using this approach, a total of 263 proteins were identified in *M. tuberculosis* and *M. bovis* BCG strains, the proteome of *M. tuberculosis* H37Rv was compared with that of *M. bovis* BCG Chicago, and twenty-five proteins differing in position or intensity were identified (Jungblut *et al.* 1999). Similarly, 137 proteins were detected in *M. tuberculosis* H37Rv culture supernatant, and 27 unique proteins were identified in *M. tuberculosis* H37Rv by comparing proteins in the culture supernatant of virulent *M. tuberculosis* H37Rv to that of attenuated *M. bovis* BCG Copenhagen (Mattow *et al.* 2003). However, recent advances in multi-dimensional liquid chromatography coupled with tandem mass spectrometry (LC/LC/MS/MS) (Washburn *et al.* 2001) have produced

a technology capable of direct analysis of the composition of protein mixtures as complex as cell lysates (Aebersold and Mann 2003).

Using this method, approximately 1,500 *S. cerevisiae* proteins were detected (Peng *et al.* 2003; Washburn *et al.* 2001). Similarly, in mycoplasma, Jaffe *et al.* detected 557 open reading frames (ORFs) in *M. pneumoniae* strain M129 by using proteogenomic mapping, the mapping of peptides detected in the cell lysate onto the uninterpreted genome (Jaffe *et al.* 2004). Here, we apply LC/LC/MS/MS to characterize the expressed proteome of *M. smegmatis,* and we report observation of 901 distinct proteins under differing growth conditions, estimate relative abundance of each protein. The protein relative abundance are correlated with codon choice and mRNA expression levels, as measured by comparison with codon adaptation indices, principal component analysis of codon frequencies and DNA microarray data. This observation is consistent with notions that either (1) prokaryotic protein expression levels are largely pre-set by codon choice, or (2) codon choice is optimized for consistency with average expression levels regardless of the mechanism of regulating expression.

RESULTS

Experimental observation of 901 proteins in the *M. smegmatis* proteome and functions of the observed proteome

Approximately 825,000 MS/MS peptide fragmentation spectra were collected and analyzed over the course of 25 LC/LC/MS/MS experiments, characterizing the proteins expressed in each of 25 samples drawn from time courses of *M. smegmatis* growing in three different media. At an estimated false-positive identification rate less than 5%, we identified a total of 901 *M. smegmatis* proteins (Figure 2.1). These identified proteins represent ~10% of the 8,968 predicted genes identified in the unfinished *M. smegmatis* genome. 94% of the proteins were detected in more than one experiment, with a few proteins (2%) detected in every one of the 25 experiments.

Each observed *M. smegmatis* protein was associated with a functional category by comparing the amino acid sequences (using BLASTP) to a database of 350,111 protein sequences from 89 fully sequenced genomes and transferring the broad-level Clusters of Orthologous Groups (COG) annotation (Tatusov *et al.* 2001) from the top-scoring homologs, where significant, to the *M. smegmatis* proteins. The broad functions of the set of 901 detected proteins are plotted in Figure 2.1. Major functions represented include energy production and conversion, amino acid transport and metabolism, translation, ribosomal structure and biogenesis, lipid transport and metabolism. Roughly 33% of the proteins could not be assigned functions in this manner (Figure 2.1).

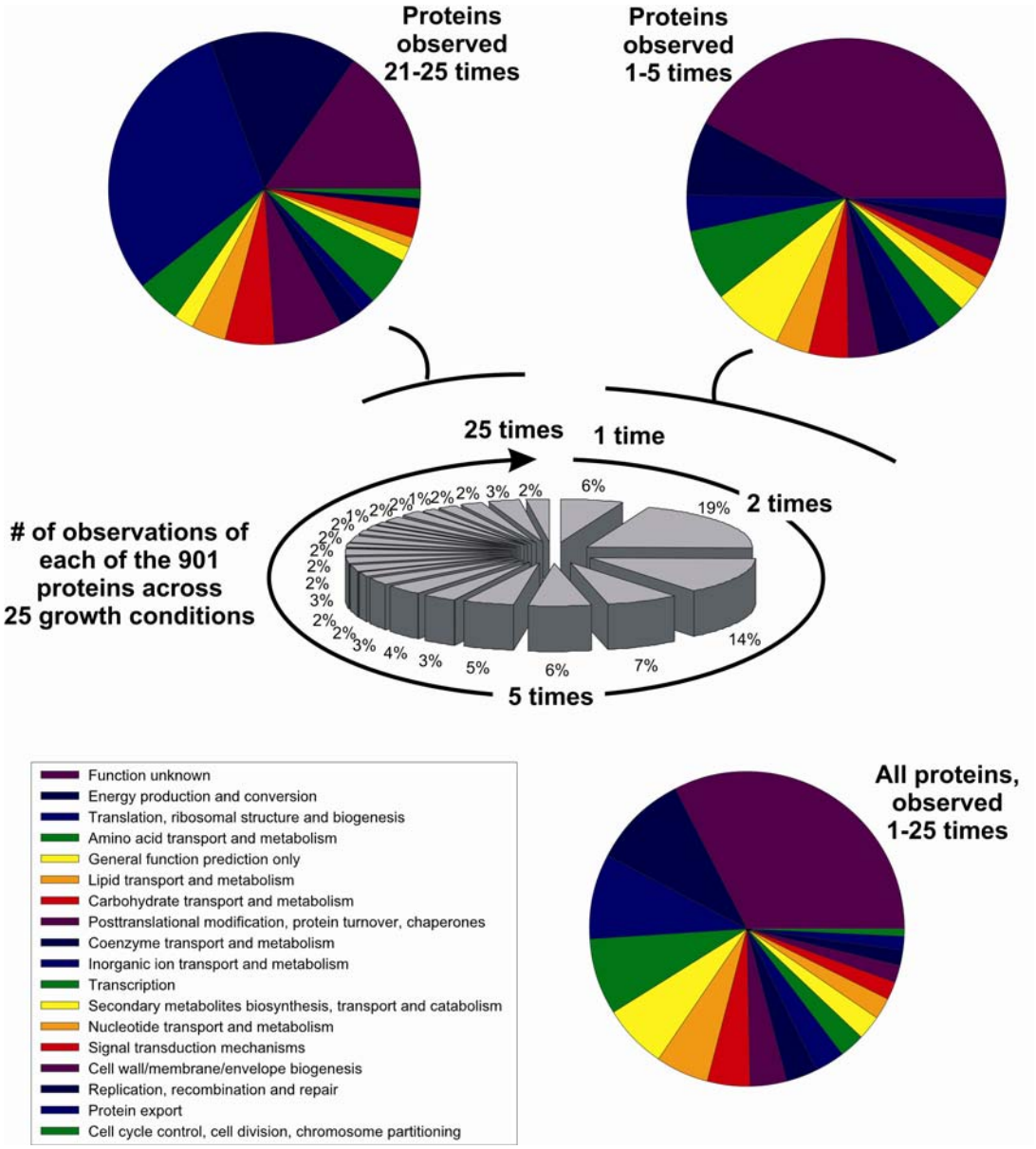


Figure 2.1 The distribution of the 901 proteins identifided across 25 LC/LC/MS/MS experiments.

The distribution of observations of each of the 901 proteins (central chart) identified across 25 LC/LC/MS/MS experiments and the associated protein functions for the complete set of proteins (bottom chart), proteins detected in one to five of the 25 experiments (top right chart), and the high abundance proteins detected in 20-25 of the 25 experiments (top left chart).

The 901 proteins were ranked by the number of observations of each protein across the 25 experiments. The set of proteins observed in 21-25 experiments, likely corresponding to highly expressed proteins, was significantly enriched for proteins involved in translation (30%), energy production (15%), small molecule transport and metabolism (18%), as well as a large fraction of uncharacterized proteins (15%). Examples of proteins in this set include peptidyl-prolyl cis-trans isomerase A (PpiA), glyceraldehyde 3-phosphate dehydrogenase (Gap), ATP synthase beta chain (AtpD), elongation factor Tu (Tuf), 60 kDa chaperonin (GroEL1), and 23 ribosomal proteins (e.g., RpsA, RplR). A number of these highly expressed proteins (3.6%) are involved in the metabolism of lipids, important components of the mycobacterial cell wall, accounting for ~60% of the cell wall weight (Ratledge 1999). Highly expressed lipid metabolism proteins included acyl carrier protein (involved in meromycolate extension), acetyl-/propionyl-coenzyme A carboxylase (AccA3) and propionyl-CoA carboxylase beta chain 5 (AccD5), enzymes responsible for creating lipid structures in the cell wall. In addition to proteins involved in lipid metabolism, the enzymes necessary for glycolysis, the tricarboxylic acid cycle, and a large number of ribosomal proteins are also highly expressed, consistent with expectation.

Focusing instead on proteins observed one to five times amongst the 25 experiments reveals a very different trend. These proteins, probably representing low expression proteins, are substantially enriched for uncharacterized proteins, especially proteins whose broad function can be approximately assigned by homology (e.g., "dehydrogenase"), but whose specific function in *M. smegmatis* is unknown. The proteins involved in translation are substantially under-represented in this set. The complete set

of protein identifications and growth phase expression levels are available as Supplemental Table 2.1.

Proteomic observation of co-expression of operon-encoded proteins

Given that operon-encoded proteins are co-transcribed and are generally co-translated, we expect that proteins in the same operon should be co-expressed across the proteomic profiling experiments. Figure 2.2 shows four such examples of proteins in the same operon coordinately expressed across the 25 experiments. In each case, proteins encoded within the operon are observed, while proteins encoded by flanking genes and those on the opposite strand are not observed. For example, Figure 2.2A shows AtpF, AtpH, AtpA, AtpG, AtpD and AtpC, subunits of the F0F1-type ATP synthase. Coexpression of these proteins reveals they are in the same operon while the neighboring genes and those encoded on the opposite strand, such as MSORF0600, 0602 and 0604 are not detected.

We see similar behavior for a uniquely mycobacterial system (Figure 2.2B): the genes that synthesize mycolic acid, the main component of the cell wall in mycobacteria and the end product in the metabolic pathway of InhA which is the primary target of mycobacterial drug isoniazid (Banerjee *et al.* 1994). KasB (3-oxoacyl-[acyl-carrier-protein] synthase2) is in the same operon as KasA (3-oxoacyl-[acyl-carrier-protein] synthase1), AcpM (acyl carrier protein), FabD (malonyl CoA-acyl carrier protein transacylase), and AccD6 (acetyl-/propionyl-CoA carboxylase). Both FabD and KasB are known components of the mycolic acid synthesis pathway (Wilson *et al.* 1999). Figure 2.2B shows that proteins encoded by the KasB operon are co-expressed, while proteins encoded by flanking genes are not.

Ribosomal proteins are often encoded in large operons to regulate ribosome synthesis (Allen *et al.* 1999). Figure 2.2C shows that the RpsJ and RpsS operons, which are components of the 11 gene S10 ribosomal protein operon, produce large and small ribosomal subunit proteins; the proteins are strongly expressed in a coordinate fashion.

Similarly, we observe an operon formed of antigen proteins (Figure 2.2D). MSORF5388 (ESAT6) and MSORF5389 (a homolog of Esat-6 protein family), identified in 17 and 19 experiments, respectively, are secreted antigens (Ratledge 1999). These two proteins are co-expressed with MSORF5386, an uncharacterized protein, suggesting that these three proteins form an operon, and by this association, may be functionally linked. Membership in the same operon argues that MSORF5386 may function with ESAT6 and MSORF5389.

Correlation between protein expression levels and mRNA expression levels of orthologous genes

We reason that the number of experiments in which a protein is observed (sampled) should roughly correlate with the protein's expression level. This assumption is based on the idea that in a LC/LC/MS/MS analysis of a cell lysate, which may consist of several hundred thousand tryptic peptides, only a subset of MS peaks are sampled for further MS/MS analysis, in total collecting ~30,000-40,000 MS/MS spectra per experiment. As the sampled peptides are typical chosen according to peak height in the MS parent spectra (here, choosing the 3 tallest peaks per MS spectrum), this introduces an intrinsic element of stochastic sampling and a bias towards high abundance proteins. Therefore, over repeated analyses, we expect to sample highly abundant proteins more

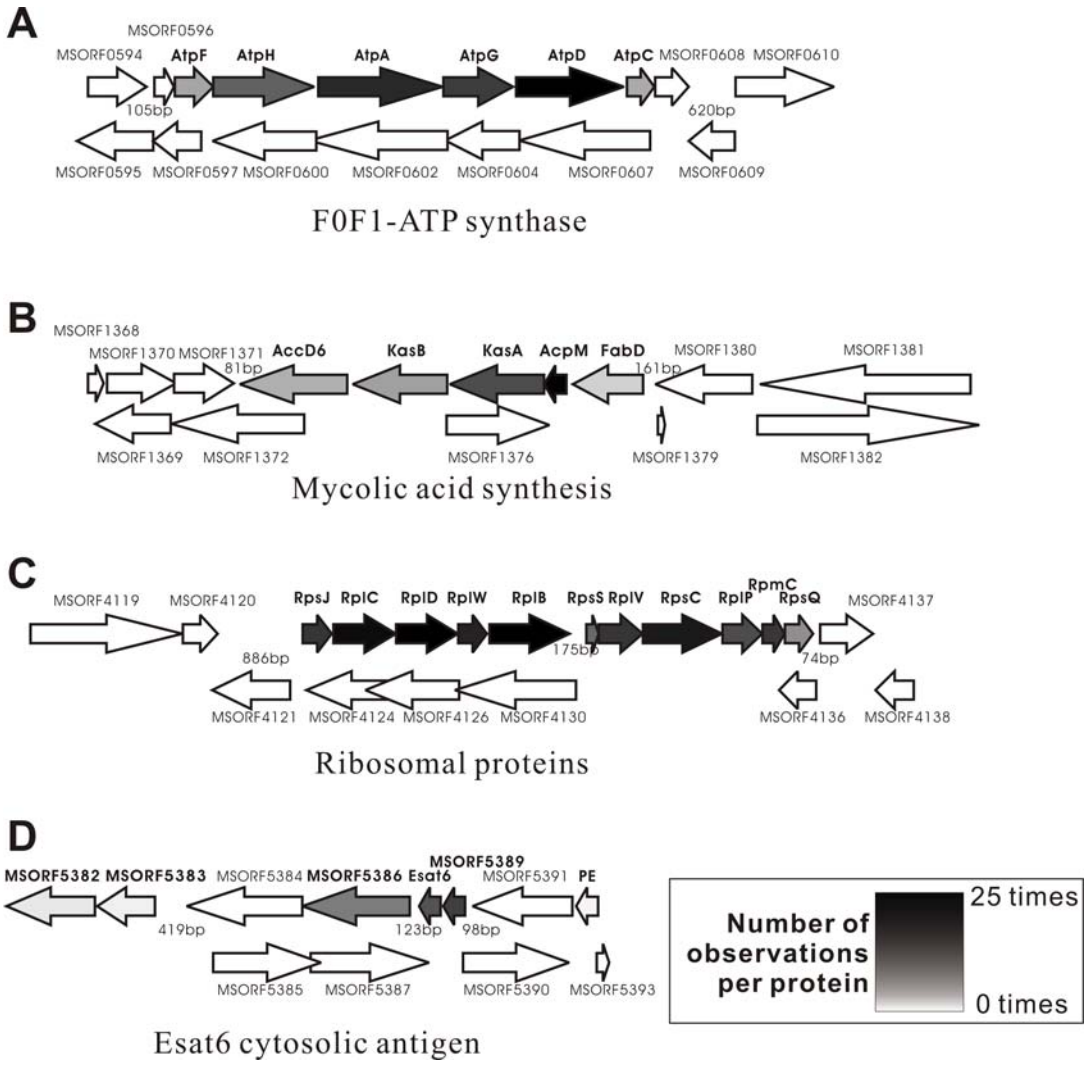


Figure 2.2 Proteins encoded in the same operon were observed to be coexpressed across the 25 experiments.

All proteins observed in LC/LC/MS/MS experiments are labeled in bold, with arrows shaded according to the number of observations.

often and lower abundant proteins less often (e.g., as recently described in (Liu et al. 2004). In the analyses presented here, rather than repeated analysis of identical samples, we have analyzed related samples, thus this trend will be conflated with the tendency for a protein to be broadly expressed, rather than just highly expressed. Nonetheless, the general trend should hold.

We wished to test our assumption that observation frequency will correlate with relative protein abundance. As a first-order test of this idea, we compared the *M. smegmatis* protein observation frequency data with mRNA expression levels, under the assumption that protein and mRNA levels should be reasonably consistent. For this test, we required an absolute measurement of mRNA expression levels, such as provided by Affymetrix style DNA microarrays. As no such data is available for *M. smegmatis* prior to completing the genome sequence, we instead used data for the orthologous genes of *E. coli* grown under roughly equivalent conditions (exponential phase growth; (Covert et al. 2004). The number of observations of *M. smegmatis* proteins correlates moderately well with the absolute abundance of their *E. coli* orthologs' mRNAs ($R^2 = 0.72$, Figure 2.3). *E. coli* orthologs of proteins observed in all 25 experiments showed high mRNA expression levels (2313 +/-1126, arbitrary units from Affymetrix array data), while *E. coli* orthologs of proteins observed in only a single experiment were significantly lower (455 +/- 557; p-value < 0.001 by t-test). An ANOVA F-test of the simple linear regression between the mRNA abundance and number of mass spectrometry observations is significant ($p<0.001$). These support the notion that the depth of sampling of a protein by mass spectrometry does indeed roughly correspond to protein abundance.

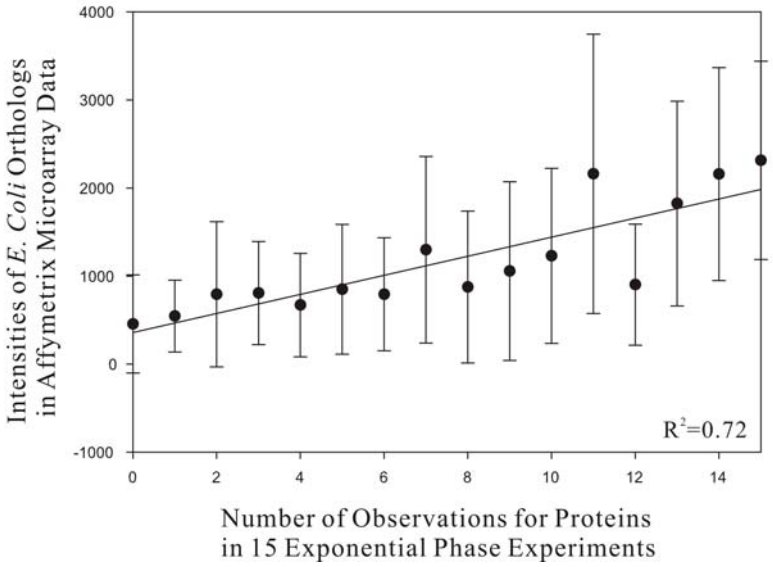


Figure 2.3 The number of observations for proteins correlates with mRNA expression level of *E. coli* orthologs.

The number of experiments in which each protein was observed in exponential phase correlates ($R^2 = 0.72$) with the corresponding mRNA expression levels, estimated from *E. coli* orthologs' measured expression (Covert *et al.* 2004), suggesting that the depth of mass spectrometry sampling provides a rough estimate of protein abundance.

Correlation between protein expression levels and the protein bias

Because the number of observations of each protein correlates at least roughly with protein expression levels, we tested if these approximate expression levels were predictable from protein amino acid sequence properties. First, we compared proteins' expression levels to their codon biases, *via* the protein codon adaptation indices (CAI) (Sharp and Li 1987), which often correlate with protein expression level (Bennetzen and Hall 1982; Futcher *et al.* 1999; Gygi *et al.* 2000; Jansen *et al.* 2003; Sharp and Li 1987). As seen in Figure 2.4A, the proteins' average CAI values are positively correlated (R^2 = 0.74) with the number of observations of each protein, suggesting the expression levels are indeed consistent with codon choice. Proteins identified in all of 25 experiments typically showed high CAI values (CAI = 0.75 +/- 0.06) —examples include proteins involved in general metabolic pathways (e.g., PpiA, Gap, AtpD, Tuf, GroEL1 and RpsA) and specific metabolic pathways, especially lipid metabolism, such as acyl carrier protein, AccA3 and AccD5. By contrast, proteins observed only in a single experiment showed considerably lower CAI values (CAI = 0.63 +/- 0.06), a significant difference in CAI under a t-test (p-value < 0.001).

To further investigate the predictability of protein expression levels, we performed principal component analysis (PCA) of *M. smegmatis* proteins' codon frequencies. PCA is a technique for summarizing the major variation in data. In PCA, the dimensionality of a data set is reduced by projecting the original data along new coordinate axes (the principal components) that capture the major trends in the data. Principal components (PC's) are linear transformations of the original sets of variables, uncorrelated and ordered, with the first few PC's containing most of the variation in the

original data set (Jolliffe 2002; Yeung and Ruzzo 2001). Performing PCA on the codon frequency vectors associated with each protein in our experiments therefore reveals the major trends in codon choice among observed *M. smegmatis* genes. While codon adaptation indices capture one aspect of codon bias, PCA should return the major trends in codon bias, regardless of whether these correlate with CAI. We tested the major PC's for correlation with the number of observations of each protein. The dominant PC (PC1) captured the background frequency of codons used by *M. smegmatis* and did not correlate with protein expression levels (data not shown). However, the second largest PC (PC2) correlated well with protein expression levels (Figure 2.4B, $R^2 = 0.84$), supporting the notion that protein expression levels are largely predictable by codon choice. PC3 and PC4 are no longer correlated with expression level. R^2 was calculated from the average values of protein expression levels for the number of observations, while ANOVA F-test was performed from the raw data. The association of both PC2 and CAI-derived codon bias with protein abundance are significant under an ANOVA F-test ($p<0.001$). Thus, codon bias, as captured by CAI or PCA, can be used to predict approximate expression levels of *M. smegmatis* proteins.

Growth-phase specific expression

The physical and metabolic changes in mycobacteria during exponential and stationary phase are largely unknown. As comparative proteomics has the potential to detect changes in the proteome, for example changes in abundance as well as post-translational modifications such as phosphorylation and glycosylation (Link *et al.* 1999), we compared the proteins expressed in exponential and stationary growth phases (Figure 2.5A). Exponential phase cultures showed a higher fraction of proteins associated with

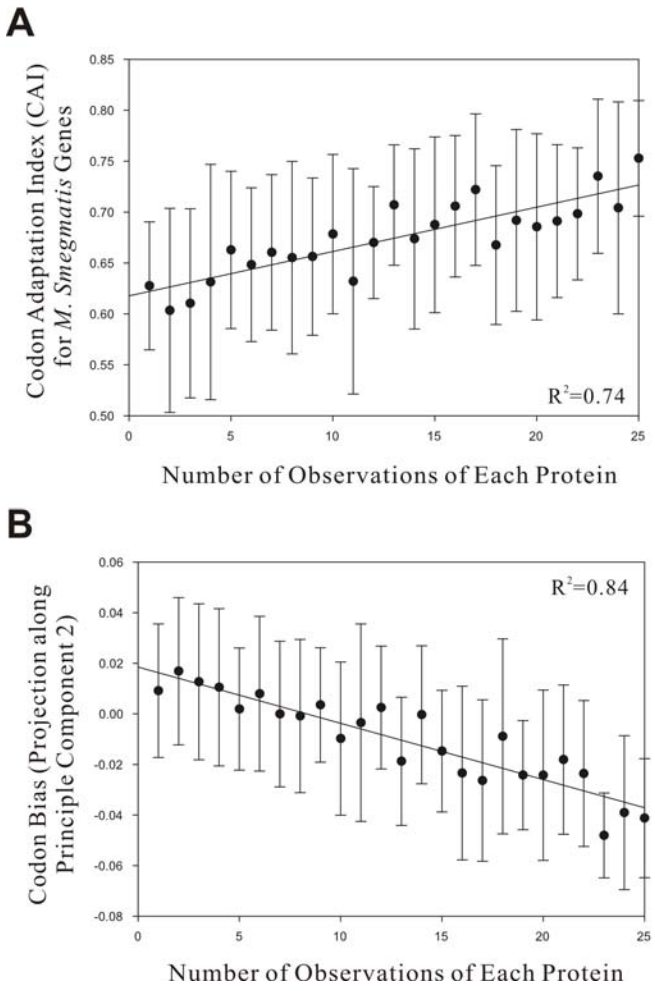


Figure 2.4 The approximate expression levels of each protein are predicted by two different measures of the codon bias.

(A) The number of experiments in which each protein was observed (estimating protein abundance) correlates with the proteins' codon adaptation indices (CAI; $R^2 = 0.74$) indicating that the proteins' average expression levels can be partially predicted from the choice of codons used for each protein. (B) The number of observations of each protein also correlates well with codon bias calculated by principal component analysis of the proteins' codon frequencies. Here, we plot the number of observations of each protein (x-axis) versus the projection of the gene's codon frequencies onto the second principal component (y-axis) (PC2; $R^2 = 0.84$), which best captures variation in codon choice between proteins with high and low expression levels. In this case, PC2 is negatively correlated with protein abundance.

active growth, such as DNA replication, recombination and repair, transcription, and translation. In contrast, stationary phase cells show higher fractions of proteins involved in competition for limited nutrients, including energy production and conversion, inorganic ion and carbohydrate transport and metabolism and posttranslational modification.

We were able to investigate more subtle differences in protein abundance between exponential and stationary phase by comparing the number of experiments in which each protein was observed in the two growth phases (Figure 2.5B and C). In this analysis, we expected to find sets of proteins overrepresented in stationary phase, presumably in order to make cells more competitive with stresses in nutrient-starved culture, and we did indeed find many such proteins. Among the proteins enriched in stationary phase cells was the transcriptional activator MtrA, which is also induced in *M. bovis* BCG upon entry into macrophages (Zahrt and Deretic 2000). Other stationary phase enriched proteins were involved in carbohydrate or fatty acid metabolism (GlgC, PrpE, LpqG, LpqY), energy metabolism (Catalase, QcrC, PdhAB, and Ndh), and sugar transport (ribose ABC transporter, GntP, β-glucanase), or amino acid transport (ProV, 3-dehydroquinase). Regulatory proteins (PhoY2) were also identified, as well as proteins of unknown function.

Among the proteins we identified as up-regulated in exponentially growing cells, a number share the same COG functions with proteins enriched in stationary phase but have different specific functions. For example, the sigma factor SigA is only detected in exponential phase—consistent with the result that SigA mediates enhanced growth of M. tuberculosis (Wu et al. 2004)—while other sigma factors are enriched in stationary phase, such as SigH. As sigma subunits of E. coli are differentially expressed according to the

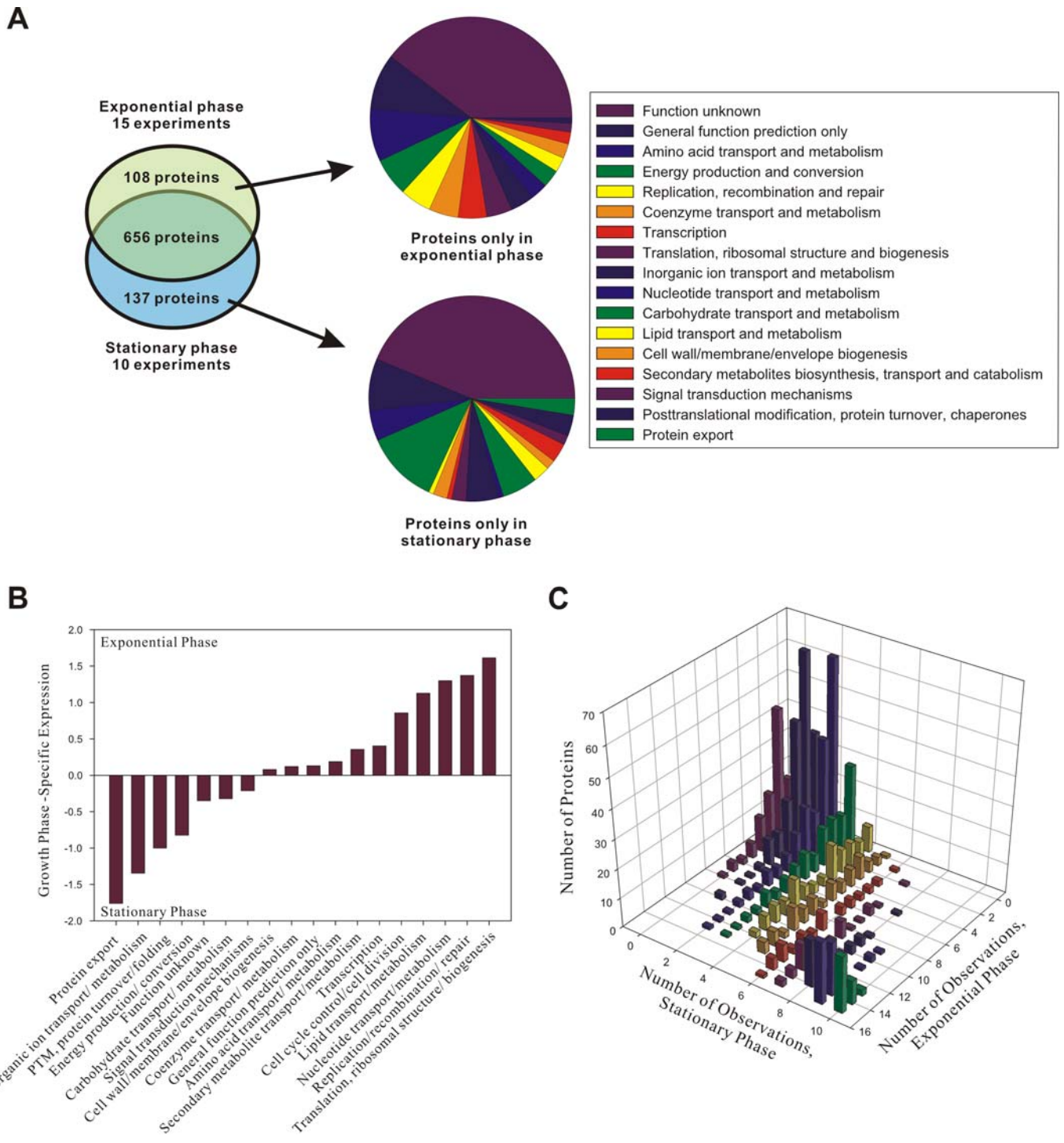

A
Exponential phase
15 experiments
108 proteins
656 proteins
137 proteins
Stationary phase
10 experiments
Proteins only in
exponential phase
Proteins only in
stationary phase
Function unknown
General function prediction only
Amino acid transport and metabolism
Energy production and conversion
Replication, recombination and repair
Coenzyme transport and metabolism
Transcription
Translation, ribosomal structure and biogenesis
Inorganic ion transport and metabolism
Nucleotide transport and metabolism
Carbohydrate transport and metabolism
Lipid transport and metabolism
Cell wall/membrane/envelope biogenesis
Secondary metabolites biosynthesis, transport and catabolism
Signal transduction mechanisms
Posttranslational modification, protein turnover, chaperones
Protein export
B
Growth Phase -Specific Expression
Exponential Phase
Stationary Phase
Protein export
Inorganic ion transport/ metabolism
PTM, protein turnover/folding
Energy production/ conversion
Function unknown
Carbohydrate transport/ metabolism
Signal transduction mechanisms
Cell wall/membrane/envelope biogenesis
Coenzyme transport/ metabolism
General function prediction only
Amino acid transport/ metabolism
Secondary metabolite transport/metabolism
Transcription
Cell cycle control/cell division
Lipid transport/metabolism
Nucleotide transport/metabolism
Replication/recombination/ repair
Translation, ribosomal structure/ biogenesis
C
Number of Proteins
Number of Observations, Stationary Phase
Number of Observations, Exponential Phase

Figure 2.5

Figure 2.5 A comparison of the proteomes of exponential and stationary phase cells.

(A) The overlap of identified proteins in exponential and stationary phases (left chart), showing the functions of proteins specific to exponential phase (top right chart) or stationary phase (bottom right chart). For clarity, proteins with uncharacterized functions are excluded from the pie charts. (B) COG functions for the proteins differentially expressed in exponential and stationary phases. For the proteins in each COG category, we plot the mean difference in expression level, calculated as mean($N_{exp,i} - 3/2*N_{stat,i}$), where $N_{exp,i}$ and $N_{stat,i}$ are the number of observations of protein i in exponential and stationary phase, respectively. Proteins of cell growth (translation, replication, etc.) are highly induced in exponential phase, while proteins of transport and energy production are highly induced in stationary phase. PTM, post-translational modification. (C) A 2D histogram plots the distribution of protein abundances and differential expression between stationary and exponential phase cells. The bulk of proteins lie on the diagonal (no differential expression), off-diagonal proteins are differentially expressed to varying degrees.

functions of the proteins they regulate (Ishihama 2000), it seems likely that SigA may be regulating housekeeping genes, while SigH regulates stationary-phase specific genes, such as stress response or transport genes (Kormanec *et al.* 2000; Thackray and Moir 2003).

In many cases the differentially expressed proteins were components of operons that were themselves differently expressed. For example, the RpsJ, RpsS, and KasB operons are up-regulated in exponential phase, while the PdhABC, Ndh, and FadE operons are up-regulated in stationary phase (Figure 2.6).

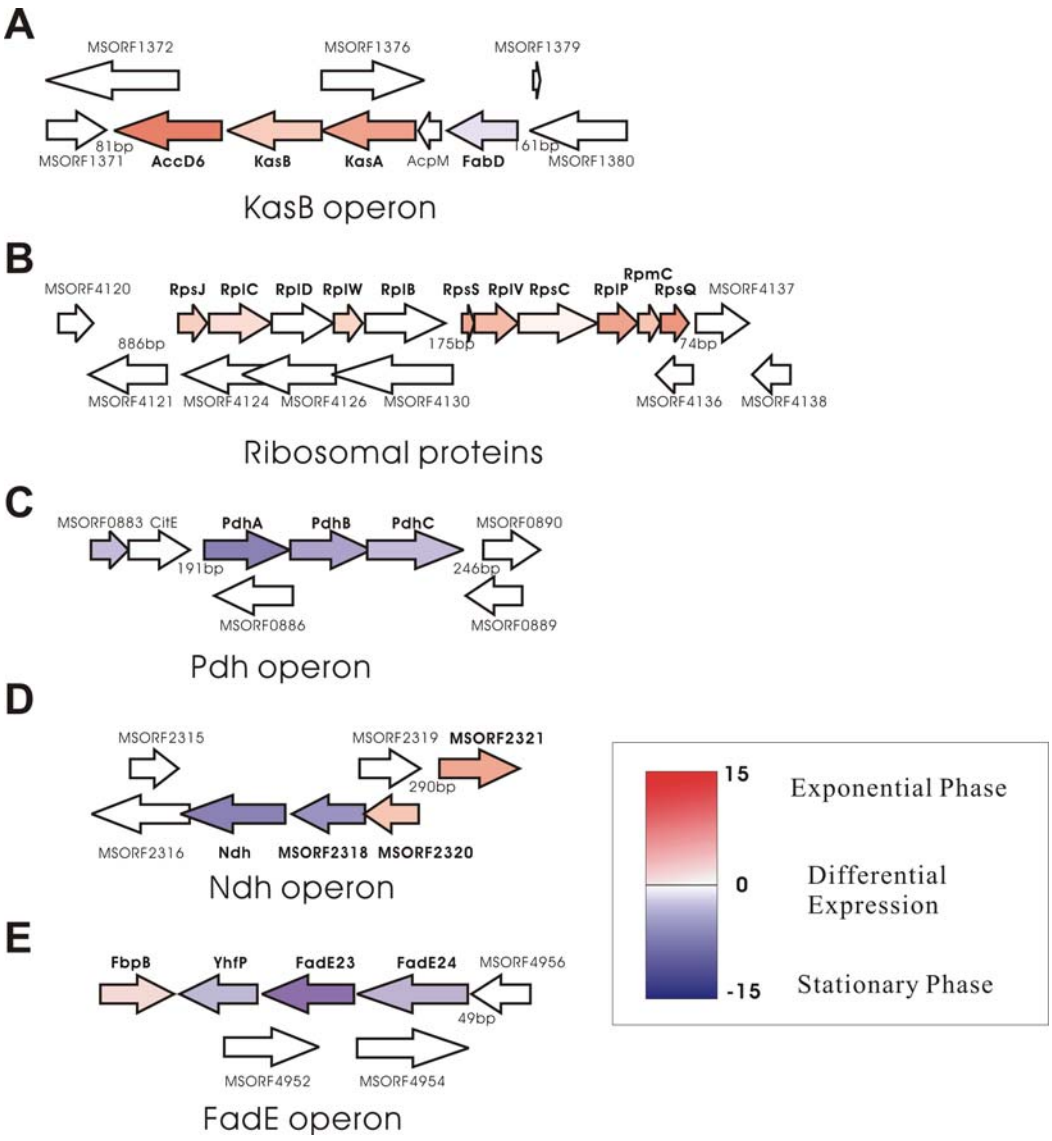


Figure 2.6 Proteins in the same operon co-express in a growth-phase specific manner.

The color scale represents the relative expression of the protein between exponential phase (red) and stationary phase (blue), calculated as (number of observations in exponential phase – number of observations in stationary phase * 3/2), where the factor of 3/2 is introduced to scale the number of stationary phase experiments to match the number of exponential phase experiments. The KasB operon (A) and RpsJ, RpsS operons (B) are up-regulated in exponential phase; the PdhABC (C), Ndh operon (D) and FadE operon (E) are up-regulated in stationary phase.

DISCUSSION

The ability to rapidly annotate genome sequences is becoming increasingly important given the current pace of genome sequencing; equally important are methods for characterizing protein dynamics on a genome-wide scale. In this study we demonstrate the suitability of coupled liquid chromatography-mass spectrometry to address both problems. Earlier analysis of mycobacterial proteomes by 2DE-MS identified roughly 300 proteins (Jungblut *et al.* 2001; Mattow *et al.* 2003). By contrast, LC/LC/MS/MS methods have proved capable of identifying ~1,500 proteins from yeast lysates (Peng *et al.* 2003; Washburn *et al.* 2001). As applied here, tandem mass-spectrometry (MS/MS) is biased in single experiments towards detection of abundant proteins (e.g., ribosomal proteins and heat shock proteins), but stochastically samples lower abundance proteins. Repeated analysis of related samples therefore leads to detection of lower expression level proteins, such as cell division proteins and transcription factors, and allows us to identify and estimate relative expression levels of 901 proteins from *M. smegmatis*. The data provide experimental annotation of predicted genes, provide measurement of relative protein abundance changes across conditions, and demonstrate the correlation of protein abundance with mRNA expression levels and codon choice.

One goal of this mass spectrometry based analysis was to provide experimental annotation of abundantly expressed proteins. In particular, a painstaking manual evaluation of each predicted gene was required for other mycobacterial genomes (e.*g.*, *M. tuberculosis* H37Rv (Cole et al. 1998a)) in order to eliminate false positive gene predictions. We expect, given the density of predicted genes and previous observations

of the performance of Glimmer2.0. on default settings (Delcher *et al.* 1999), that a significant fraction of the predicted genes are false positives. Assuming similar gene density to *M. tuberculosis* or *M. bovis* BCG, we expect ~6,300 genes in *M. smegmatis* (David Graham, personal communication), suggesting that ~30% of the 8,968 predicted genes are false positives. For example, genes predicted on the opposing strand to expressed proteins within the operons of Figure 2.2, such as MSORF1376 in the KasB operon, may be candidate false positive predicted genes.

Note that unlike other applications, for the purpose of constructing a mass spectrometry reference database, we desire to minimize the false negative gene identification rate, even at the expense of increasing the false positive gene identification rate, to ensure that all representative sequences are present. These gene prediction errors are clearly undesirable for other purposes, but for mass spectrometry serve to minimize the false negative protein identification rate in the mass spectrometry experiment. The extreme case of this strategy is to compare the MS/MS spectral data against raw, uninterpreted genomic sequence data (Arthur and Wilkins 2003), even without predicting genes (Jaffe *et al.* 2004). In this mode, the false positive gene prediction rate is maximal; however, this minimizes the false negative protein identification rate, allowing previously unrecognized proteins to be identified. Proteomics has previously helped in this manner: for example, of 263 proteins detected in *M. tuberculosis* H37Rv by 2DE-MS, 6 were not previously predicted (Jungblut *et al.* 2001). To aid in such future annotation studies, all raw mass spectrometry data have been deposited into the public domain in the Open Proteomics Database (Prince *et al.* 2004). However, v*ia* this mass spectrometric approach, we are able to rapidly validate a significant subset of the predicted genes. The fraction of the proteome we identified also provides a framework

for understanding the activity of pathogenic *M. tuberculosis*: in total, 709 (78.7%) of the 901 proteins we detected have homologs in *M. tuberculosis* H37Rv, compared to only 47.5% for the entire set of 8,968 predicted *M. smegmatis* proteins, or to 68% if we assume. Thus, the expressed *M. smegmatis* proteome is enriched for proteins of relevance to *M. tuberculosis*.

Beyond experimentally annotating expressed proteins, the frequency of observation of each protein gives an estimate of protein abundance. We took advantage of this data to estimate the correlation between codon bias and protein abundance. Such a correlation has previously been observed for high abundant proteins, for example, in two-dimensional gel electrophoresis analysis of the yeast proteome (Futcher *et al.* 1999; Gygi *et al.* 2000). In our hands, the *M. smegmatis* protein expression levels were predictable from the genes' codon biases as calculated by codon adaptation indices, as well as by principal component analysis. Our data supports two models of prokaryotic protein regulation. In the first, the condition under which a protein would be expressed is set by the action of transcription factors and regulatory proteins, but the quantity of the expressed protein is "pre-set" by its codon bias. Under this view, regulation of a prokaryotic protein is simplified to deciding when to express it at a given time, with the actual amount of the protein synthesized being optimized over evolutionary time. A perhaps more plausible, equally supported explanation is that codon choice has been optimized to be non-limiting. That is, protein abundance is set in some fashion (e.g., translation initiation, promoter strength, *etc.*) and may be dynamically regulated, but rarely exceeds a characteristic expression level. Codons might then face selective evolutionary pressure to optimize such that the codon choice is non-limiting for the normal range of each protein's expression. This scenario would still produce a correlation

between codon bias and average protein abundance. We note that these expression measurements could in principle be made more accurate through the use of isotope labeling approaches, such as the use of metabolic labeling of stable-isotope tags or ICAT (Gygi *et al.* 1999) for quantitative protein expression profiling (Flory *et al.* 2002), as used to quantitate 280 *M. tuberculosis* proteins (Schmidt *et al.* 2004).

Numerous physiological and metabolic changes happen during the transition from exponential phase to stationary phase, and we capture some of these changes in the differential expression analysis. Proteins involved in the pathways of DNA replication, recombination and repair, transcription, and translation show a higher fraction of proteins associated with active growth. In accord with our results, UmaA1 (lipid biosynthesis) and ParA (functions in chromosome partitioning), which we observe enriched in exponential phase, have been reported to be down-regulated in *M. tuberculosis* in nutrient starved media (Betts *et al.* 2002).

When cells enter stationary phase, many growth-related genes are down-regulated; instead, the stationary-phase-specific genes are expressed, such as proteins involved in energy production and conversion, post translation modification and carbohydrate transport and metabolism (Figure 2.5A). The PdhABC proteins are up-regulated in stationary phase, consistent with their up-regulation in *M. tuberculosis* starved for nutrients (Betts *et al.* 2002). We speculate Mpt53, an immunogenic protein induced only in stationary phase, the transcription factor MtrA, and several proteases (ClpP, MSORF2945, and MSORF8009, a zinc metalloprotease) up-regulated in stationary phase, may be important for stress response and therefore might also be important to the virulence of pathogenic mycobacteria. Finally, we noticed the

enrichment of numorous uncharacterized proteins in stationary phase. Although their exact function remains unknown, these proteins are likely to be stress response proteins, heat-shock proteins, or other proteins involved in coping with the nutrient limitation and high cell density in this phase of growth, suggesting that future experiments directly monitoring the stress-response of mycobacteria may be useful in more precisely determining their functions.

MATERIALS AND METHODS

Initial identification of *M. smegmatis* protein coding sequences

The partially complete genomic sequence for *M. smegmatis* mc^{2}155 was obtained from The Institute for Genomic Research (TIGR) through the website at http://www.tigr.org in the form of 15 sequence contigs, and protein encoding genes were predicted by searching with the gene-finding program Glimmer 2.0 using default settings, which minimize the false negative identification rate of coding sequences but increase the false positive rate (Delcher *et al.* 1999). A total of 8,968 genes were predicted and assembled into a searchable database for interpreting mass spectrometry peptide fragmentation spectra data using the program BioWorks 3.1 (ThermoFinnigan). The predicted genes were functionally annotated by comparing the predicted amino acid sequences against those of 350,111 protein sequences from 89 fully sequenced genomes (downloaded from the Entrez genome database) using the program BLASTP and default parameters, and selecting the protein functional annotation of the top BLAST match when it surpassed a BLAST expectation value threshold of 1e-6. A broad functional category was also assigned to each protein by selecting the COG database annotation (Tatusov *et al.* 2001) associated with the top BLAST hit, when significant. For clarity in all discussion, COG category N ("cell mobility") was labeled "protein export", as *M. smegmatis* is non-mobile and the *M. smegmatis* proteins in this category are predominantly involved in protein export and stress adaptation.

Growth of *M. smegmatis* mc^{2}155 and preparation of cell lysates

M. smegmatis mc^2155 bacteria were grown with agitation at 37°C in minimal medium (Brosch *et al.* 2001; Chacon *et al.* 2002) and rich media (Middlebrook7H9 medium and Luria broth (LB) medium) (Jacobs *et al.* 1991), collecting samples for 16, 6 and 3 time points from Middlebrook7H9 medium, LB, and minimal medium respectively. Samples were centrifuged at 12,000g for 30 min, suspended in ice-cold lysis buffer (25mM Tris HCL pH7.5, 2.5mM DTT, 1.0mM EDTA, 0.02%(w/v) Brij35, 1X Calbiochem Protease Inhibitor Cocktail Set I (CPICSI)) (1ml/g cell pellet), and disrupted by bead-beating with 1mm glass beads (Parish 2001; Primm *et al.* 2000). Cell lysates were clarified by centrifugation at 20,000g for 30 min, with typical protein concentrations of 10mg/ml.

Preparation and LC/LC/MS/MS analysis of *M. smegmatis* peptides

M. smegmatis soluble protein extracts were diluted in digestion buffer (50mM Tris HCL pH8.0, 1.0M Urea, 2.0mM $CaCl_2$), denatured at 95°C for 15 min, and digested with sequencing grade trypsin (Sigma) at 37°C for 20 h. Tryptic peptide mixtures were separated by automated two dimensional-high performance liquid chromatography. Chromatography was performed at 2µL/min with all buffers acidified with 0.1% formic acid. Chromatography salt step fractions were eluted from a strong cation exchange column (SCX) with a continuous 5% acetonitrile (ACN) background and 10 minute salt bumps of 0, 20, 40, 60, 80, 100, 150, 200, 300, 500, and 900 mM ammonium chloride (Figure 2.7). Each salt bump was eluted directly onto a reverse phase C18 column and washed free of salt. Reverse phase chromatography was run in a 60 minute gradient from 5% to 45% ACN, then purged at 95% ACN. Peptides were analyzed online with electrospray ionization (ESI) ion trap mass spectrometry (MS) (Link *et al.* 1999; Washburn *et al.* 2001) using a ThermoFinnigan Surveyor/DecaXP+ instrument. In each

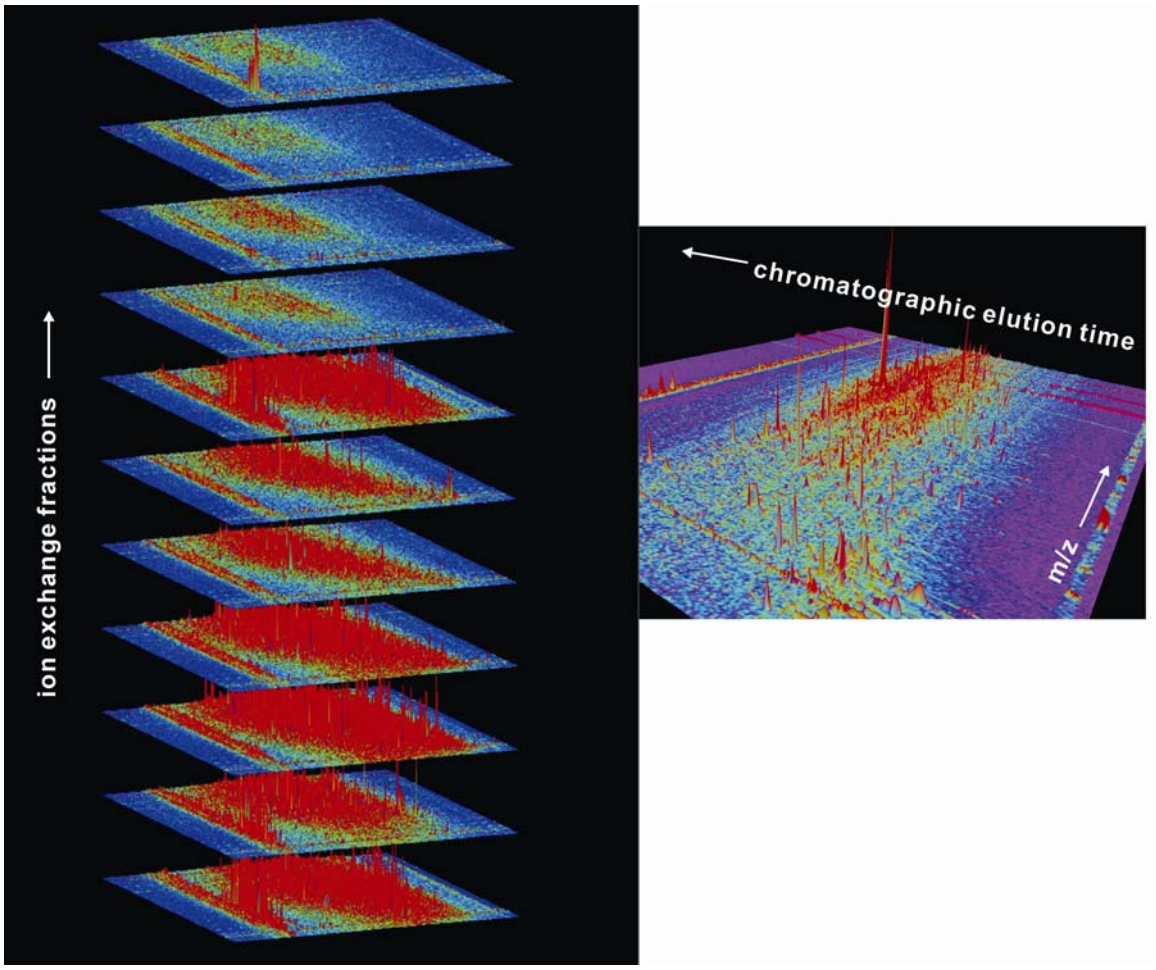


Figure 2.7 The MS spectra from 11 salt steps.

The MS spectra from 11 salt steps (left plot), the salt concentration increases from the bottom to top. Each peak corresponds to one peptide.

MS spectra, the 3 tallest individual peaks, corresponding to peptides, were fragmented by collision-induced dissociation with helium gas to produce MS/MS spectra.

Estimating the mRNA expression levels

E. coli orthologs of *M. smegmatis* proteins were identified by using BLASTP under default settings and the bi-directional best hit method (Overbeek *et al.* 1999), only accepting hits with BLAST E-values less than 1e-10. By this approach, 1,250 *M. smegmatis* predicted genes have *E. coli* K-12 orthologs. The mRNA abundance of each of these genes was calculated as the average abundance across three replicate Affymetrix DNA microarrays from (Covert *et al.* 2004), representing the mRNA abundance of *E. coli* K-12 MG1655 cells growing in M9 media in exponential phase.

Calculation of codon bias

The codon adaptation index (CAI) of each *M. smegmatis* open reading frame was calculated as in (Sharp and Li 1987). A second estimate of codon bias was generated by performing principal component analysis (Jolliffe 2002; Yeung and Ruzzo 2001) of the codon frequencies associated with each of the 901 observed *M. smegmatis* proteins. Using a Perl program, a vector was created for each gene composed of the usage frequencies of each of the 61 possible amino-acid encoding codons. Principal component analysis was performed on the resulting frequency vectors using the program Cluster (Eisen *et al.* 1998). Projection of each gene's frequency vector onto each of the principal components associated a set of numerical indices with each gene describing the major trends in the gene's codon frequencies.

An error model for mass spectrometry proteomics data

Proteins were identified from the resulting peptide MS/MS fragmentation spectra by searching against the custom *M. smegmatis* predicted protein database using the program BioWorks 3.1 (Xcorr >= 2.5) and filtering the results with selection criteria that minimize the false positive identification rate to less than 5%. Calculation of false positive identification rate was based on the following procedure: The amino acid sequences of each predicted protein encoded in the *M. smegmatis* genome were shuffled, thereby preserving the length and amino acid frequency distribution of each protein but not the amino acid order, and fragmentation spectra from the 25 LC/LC/MS/MS experiments were analyzed against the shuffled database by using the program BioWorks3.2. For each protein identified, we calculated the sum of the BioWorks protein scores across the 25 experiments using either the correct or shuffled version of the database (Figure 2.8). At a protein score threshold corresponding to a 5% false-positive identification rate in the shuffled database, a total of 899 proteins were found. A similar analysis using DTASelect (Tabb 2003) identified 426 proteins versus 2 in the shuffled database, at a false positive rate of 0.5%. By this analysis, DTASelect therefore gives a lower false positive identification rate, with a higher false negative identification rate. Using either criterion gave a total of 901 *M. smegmatis* proteins identified at a false positive rate of ~5%.

The mass spectrometry raw data from this study have been deposited in the Open Proteomics Database http://bioinformatics.icmb.utexas.edu/OPD, under accession nos. opd00007_MYCSM–opd00031_MYCSM.

Supplemental material is available online at http://polaris.icmb.utexas.edu/people/rong/dissertation.

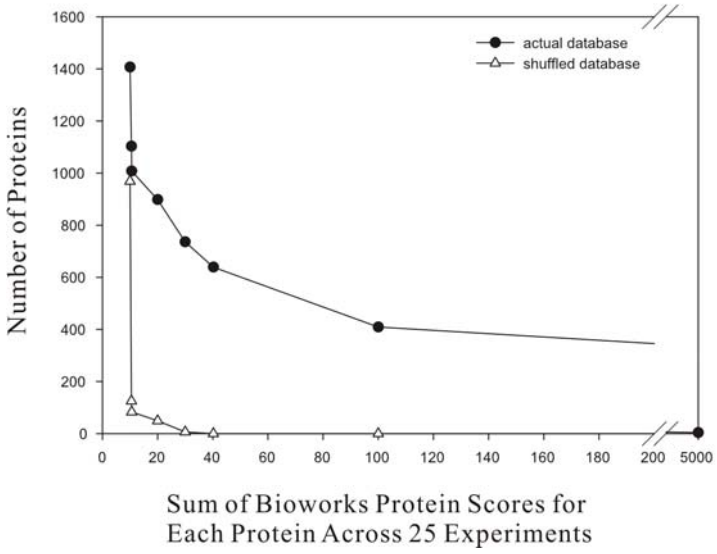


Figure 2.8 Estimating the protein identification error rate.

The false positive protein identification rate (triangles; estimated by comparing MS/MS spectra against a mass spectrometry reference database of shuffled *M. smegmatis* proteins) is plotted as a function of total BioWorks score for each protein. At a false positive identification rate of ~5% (corresponding to proteins with total scores >20 across the 25 experiments), 899 proteins are identified (circles).

REFERENCES

Aebersold, R. and M. Mann. 2003. Mass spectrometry-based proteomics. *Nature* **422:** 198-207.

Allen, T., P. Shen, L. Samsel, R. Liu, L. Lindahl, and J.M. Zengel. 1999. Phylogenetic analysis of L4-mediated autogenous control of the S10 ribosomal protein operon. *J Bacteriol* **181:** 6124-6132.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Arthur, J.W. and M.R. Wilkins. 2003. Using proteomics to mine genome sequences. *Journal of proteome research* **3:** 393-402.

Banerjee, A., E. Dubnau, A. Quemard, V. Balasubramanian, K.S. Um, T. Wilson, D. Collins, G. de Lisle, and W.R. Jacobs, Jr. 1994. inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis. *Science* **263:** 227-230.

Bennetzen, J.L. and B.D. Hall. 1982. Codon selection in yeast. *J Biol Chem* **257:** 3026-3031.

Betts, J.C., P.T. Lukey, L.C. Robb, R.A. McAdam, and K. Duncan. 2002. Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling. *Mol Microbiol* **43:** 717-731.

Boshoff, H.I., T.G. Myers, B.R. Copp, M.R. McNeil, M.A. Wilson, and C.E. Barry, 3rd. 2004. The transcriptional responses of Mycobacterium tuberculosis to inhibitors of metabolism: novel insights into drug mechanisms of action. *J Biol Chem* **279:** 40174-40184.

Brosch, R., A.S. Pym, S.V. Gordon, and S.T. Cole. 2001. The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol* **9:** 452-458.

Chacon, O., Z. Feng, N.B. Harris, N.E. Caceres, L.G. Adams, and R.G. Barletta. 2002. Mycobacterium smegmatis D-Alanine Racemase Mutants Are Not Dependent on D-Alanine for Growth. *Antimicrob Agents Chemother* **46:** 47-54.

Cole, S.T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry, 3rd, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M.A. Quail, M.A. Rajandream, J. Rogers, S.

Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J.E. Sulston, K. Taylor, S. Whitehead, and B.G. Barrell. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393:** 537-544.

Corbett, E.L., C.J. Watt, N. Walker, D. Maher, B.G. Williams, M.C. Raviglione, and C. Dye. 2003. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med* **163:** 1009-1021.

Covert, M.W., E.M. Knight, J.L. Reed, M.J. Herrgard, and B.O. Palsson. 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429:** 92-96.

de Hoog, C.L. and M. Mann. 2004. Proteomics. *Annu Rev Genomics Hum Genet* **5:** 267-293.

Delcher, A.L., D. Harmon, S. Kasif, O. White, and S.L. Salzberg. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27:** 4636-4641.

Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95:** 14863-14868.

Eisenberg, D., E.M. Marcotte, I. Xenarios, and T.O. Yeates. 2000. Protein function in the post-genomic era. *Nature* **405:** 823-826.

Flory, M.R., T.J. Griffin, D. Martin, and R. Aebersold. 2002. Advances in quantitative proteomics using stable isotope tags. *Trends Biotechnol* **20:** S23-29.

Futcher, B., G.I. Latter, P. Monardo, C.S. McLaughlin, and J.I. Garrels. 1999. A sampling of the yeast proteome. *Mol Cell Biol* **19:** 7357-7368.

Gerber, S.A., J. Rush, O. Stemman, M.W. Kirschner, and S.P. Gygi. 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **100:** 6940-6945.

Gygi, S.P., G.L. Corthals, Y. Zhang, Y. Rochon, and R. Aebersold. 2000. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A* **97:** 9390-9395.

Gygi, S.P., B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, and R. Aebersold. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17:** 994-999.

Ishihama, A. 2000. Functional modulation of Escherichia coli RNA polymerase. *Annu Rev Microbiol* **54:** 499-518.

Ishihama, Y., Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, and M. Mann. 2005. Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. *Mol Cell Proteomics* **4:** 1265-1272.

Jacobs, W.R., Jr., G.V. Kalpana, J.D. Cirillo, L. Pascopella, S.B. Snapper, R.A. Udani, W. Jones, R.G. Barletta, and B.R. Bloom. 1991. Genetic systems for mycobacteria. *Methods Enzymol* **204:** 537-555.

Jaffe, J.D., H.C. Berg, and G.M. Church. 2004. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4:** 59-77.

Jansen, R., H.J. Bussemaker, and M. Gerstein. 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res* **31:** 2242-2251.

Jolliffe, I.T. 2002. *Principal component analysis*. Springer.

Jungblut, P.R., E.C. Muller, J. Mattow, and S.H. Kaufmann. 2001. Proteomics reveals open reading frames in Mycobacterium tuberculosis H37Rv not predicted by genomics. *Infect Immun* **69:** 5905-5907.

Jungblut, P.R., U.E. Schaible, H.J. Mollenkopf, U. Zimny-Arndt, B. Raupach, J. Mattow, P. Halada, S. Lamer, K. Hagens, and S.H. Kaufmann. 1999. Comparative proteome analysis of Mycobacterium tuberculosis and Mycobacterium bovis BCG strains: towards functional genomics of microbial pathogens. *Mol Microbiol* **33:** 1103-1117.

Kormanec, J., B. Sevcikova, N. Halgasova, R. Knirschova, and B. Rezuchova. 2000. Identification and transcriptional characterization of the gene encoding the stress-response sigma factor sigma(H) in streptomyces coelicolor A3(2). *FEMS Microbiol Lett* **189:** 31-38.

Lee, I., S.V. Date, A.T. Adai, and E.M. Marcotte. 2004. A probabilistic functional network of yeast genes. *Science* **306:** 1555-1558.

Link, A.J., J. Eng, D.M. Schieltz, E. Carmack, G.J. Mize, D.R. Morris, B.M. Garvik, and J.R. Yates, 3rd. 1999. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **17:** 676-682.

Liu, H., R.G. Sadygov, and J.R. Yates, 3rd. 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76:** 4193-4201.

Lu, P., C. Vogel, and E.M.E. Marcotte. 2005. An estimate of relative contributions of transcriptional and translational regulation by absolute protein expression profiling. *submitted*.

Marcotte, E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* **402:** 83-86.

Mattow, J., U.E. Schaible, F. Schmidt, K. Hagens, F. Siejak, G. Brestrich, G. Haeselbarth, E.C. Muller, P.R. Jungblut, and S.H. Kaufmann. 2003. Comparative proteome analysis of culture supernatant proteins from virulent Mycobacterium tuberculosis H37Rv and attenuated M. bovis BCG Copenhagen. *Electrophoresis* **24:** 3405-3420.

Oda, Y., K. Huang, F.R. Cross, D. Cowburn, and B.T. Chait. 1999. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci U S A* **96:** 6591-6596.

Ong, S.E., B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, and M. Mann. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1:** 376-386.

Ong, S.E. and M. Mann. 2005. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* **1:** 252-262.

Overbeek, R., M. Fonstein, M. D'Souza, G.D. Pusch, and N. Maltsev. 1999. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96:** 2896-2901.

Pandey, A. and M. Mann. 2000. Proteomics to study genes and genomes. *Nature* **405:** 837-846.

Parish, T. and N.G. Stoker. 2001. *Mycobacterium tuberculosis Protocols.* Humana press.

Peng, J., J.E. Elias, C.C. Thoreen, L.J. Licklider, and S.P. Gygi. 2003. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* **2:** 43-50.

Primm, T.P., S.J. Andersen, V. Mizrahi, D. Avarbock, H. Rubin, and C.E. Barry, 3rd. 2000. The stringent response of Mycobacterium tuberculosis is required for long-term survival. *J Bacteriol* **182:** 4889-4898.

Prince, J.T., M.W. Carlson, R. Wang, P. Lu, and E.M. Marcotte. 2004. The need for a public proteomics repository. *Nat Biotechnol* **22:** 471-472.

Rapaport, E., A. Levina, V. Metelev, and P.C. Zamecnik. 1996. Antimycobacterial activities of antisense oligodeoxynucleotide phosphorothioates in drug-resistant strains. *Proc Natl Acad Sci U S A* **93:** 709-713.

Ratledge, C.a.D., J. 1999. *Mycobacteria molecular biology and virulence*. Blackwell science.

Schmidt, F., S. Donahoe, K. Hagens, J. Mattow, U.E. Schaible, S.H. Kaufmann, R. Aebersold, and P.R. Jungblut. 2004. Complementary analysis of the Mycobacterium tuberculosis proteome by two-dimensional electrophoresis and isotope-coded affinity tag technology. *Mol Cell Proteomics* **3:** 24-42.

Sharp, P.M. and W.H. Li. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15:** 1281-1295.

Tabb, D.L., McDonald, W.H., and Yates III, J.R. 2002. DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. 1: 21–26.

Talaat, A.M., S.T. Howard, W.t. Hale, R. Lyons, H. Garner, and S.A. Johnston. 2002. Genomic DNA standards for gene expression profiling in Mycobacterium tuberculosis. *Nucleic Acids Res* **30:** e104.

Tatusov, R.L., D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, and E.V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29:** 22-28.

Thackray, P.D. and A. Moir. 2003. SigM, an extracytoplasmic function sigma factor of Bacillus subtilis, is activated in response to cell wall antibiotics, ethanol, heat, acid, and superoxide stress. *J Bacteriol* **185:** 3491-3498.

Washburn, M.P., D. Wolters, and J.R. Yates, 3rd. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19:** 242-247.

Washburn, M.P. and J.R. Yates, 3rd. 2000. Analysis of the microbial proteome. *Curr Opin Microbiol* **3:** 292-297.

Wilson, M., J. DeRisi, H.H. Kristensen, P. Imboden, S. Rane, P.O. Brown, and G.K. Schoolnik. 1999. Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization. *Proc Natl Acad Sci U S A* **96:** 12833-12838.

Wu, S., S.T. Howard, D.L. Lakey, A. Kipnis, B. Samten, H. Safi, V. Gruppo, B. Wizel, H. Shams, R.J. Basaraba, I.M. Orme, and P.F. Barnes. 2004. The principal sigma factor sigA mediates enhanced growth of Mycobacterium tuberculosis in vivo. *Mol Microbiol* **51:** 1551-1562.

Yeung, K.Y. and W.L. Ruzzo. 2001. Principal component analysis for clustering gene expression data. *Bioinformatics* **17:** 763-774.

Zahrt, T.C. and V. Deretic. 2000. An essential two-component signal transduction system in Mycobacterium tuberculosis. *J Bacteriol* **182:** 3832-3838.

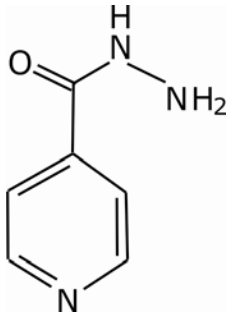
Zhang, Y. 2005. The magic bullets and tuberculosis drug targets. *Annu Rev Pharmacol Toxicol* **45:** 529-564.

Chapter 3. Proteomic analysis of drug treated *Mycobacterium smegmatis*

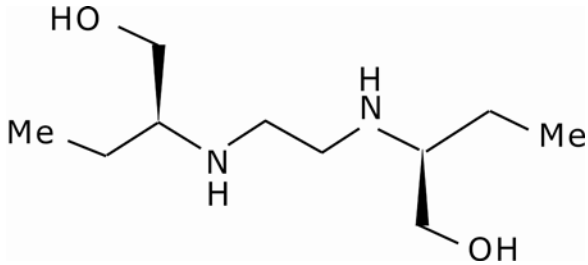
Tuberculosis (TB) is one of the most prevalent infectious diseases in the world, claiming 2 million lives every year (Corbett et al. 2003). Lengthy chemotherapy and emergence of drug-resistant strains pose significant problems for effective control. In order to shorten the duration of TB therapy, novel drugs are needed against the causative agent of TB, *Mycobacterium tuberculosis.* Understanding the biological action of the currently used drugs should help in the development of new drugs. Isoniazid (INH), ethambutol (EMB), pyrazinamide (PZA) (Figure 3.1) are the front-line drugs currently used to treat TB (Zhang 2005). INH selectively inhibits the synthesis of mycolic acids, the main component of the waxy cell wall in mycobacteria (Mdluli et al. 1998). Both enoyl-acyl carrier protein reductase (InhA) and 3-oxoacyl-[acyl-carrier-protein] synthase 2 (KasB) have been proposed as the primary targets of INH, which are involved in the type II fatty acid synthase (FAS II) system of mycobacteria for full-length extension of the meromycolate chain (Takayama et al. 2005). The target gene of EMB is arabinosyl transferase (EmbABC), which is responsible for the synthesis of the arabinan portion of arabinogalactan (AG) (Takayama and Kilburn 1989) and lipoarabinomannan (LAM) (Deng et al. 1995; Mikusova et al. 1995) in mycobacterial cell walls. PZA disrupts cell membrane function and depletes energy transport (Zhang 2005; Zhang and Mitchison 2003). Relatively little is known about the targets of PZA, although fatty acid synthase (Fas) has been proposed as a potential target (Zimhony et al. 2000).

Proteomics is capable of investigating the systematic response of mycobacteria in a reasonably comprehensive manner (Washburn and Yates 2000). Two-dimensional gel

(A)



(B)



(C)

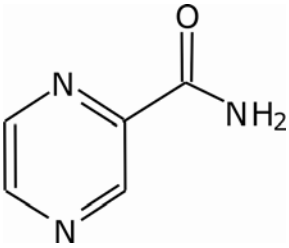


Figure 3.1 Structures of anti-TB drugs.

Structure of anti-TB drugs (A) isoniazide (INH), (B) ethambutol (EMB), and (C) pyrazinamide (PZA).

electrophoresis coupled with mass spectrometry (2DE-MS) was a popular analytical method used in proteomics study. A total of 263 proteins were identified in *M. tubeculosis* and *M. bovis* BCG strains (Jungblut et al. 1999). 108 proteins (Schmidt et al. 2004) and 105 membrane proteins (Sinha et al. 2005) were identified in *M. tuberculosis*. MudPIT, multi-dimensional protein identification technology, combining multi-dimensional chromatographic separation and mass spectrometric detection (LC/LC/MS/MS) (Washburn et al. 2001), has been proven to be very efficient for analysis of highly complex protein mixtures as cell lysates (Aebersold and Mann 2003). This technology can be used to do high-throughput analyses while retaining high sensitivity and highly reproducible performance. Using this technique, 1044 proteins were identified from three subcellular compartments in *M. tuberculosis* (Mawuenyega et al. 2005) and 901 proteins were identified from the *M. smegmatis* cell lysate (Wang et al. 2005).

Proteomics is rapidly evolving from analysis of global expression patterns toward characterization of differential expression. Protein profiles after exposure to drugs make it possible to classify compounds of unknown mechanism of action into similar groups, thereby elucidating the site of mechanism of action (Takayama and Kilburn 1989). This idea was explored in our work to study the dynamically changing proteome, to compare proteomic fingerprints of drugs, and to identify markers of drug action. The fast-growing *Mycobacterium smegmatis* is a non-pathogenic model system for studying the cellular processes of mycobacteria, such as the related pathogenic species *M. tuberculosis*. It has been used to study alterations in protein expression in anti-TB drug treatments (Takayama and Kilburn 1989; Zimhony et al. 2000). *M. smegmatis* is sensitive to INH, EMB, and PZA's analog 5-Cl-PZA (Zimhony et al. 2000), but resistant to PZA. Using

LC/LC/MS/MS we analyzed and quantified the *M. smegmatis* protein profiles in response to INH, EMB, and 5-Cl-PZA, which allowed us to identify proteins and families of proteins that show altered expression levels and elucidate the drugs' systematic effects on mycobacterial cells.

RESULTS

The growth of *M. smegmatis* cells are inhibited by anti-TB drugs

To investigate the metabolic changes, protein expression was examined throughout a time course after adding the anti-TB drugs. The *M. smegmatis* mc^2155 cells were grown at 37˚C to early log phase, anti-TB drugs were added, and cells were grown with agitation at 37˚C. At initial time points, cells appear normal, with increasing time cell growth slows. In the presence of 4 μg INH per ml, the growth rate of *M. smegmatis* was reduced, and at 8 μg/ml, INH completely inhibited growth. In EMB treatment, when EMB concentration is 1 μg/ml, the cell growth was reduced. Cell growth stops when EMB reaches 4 μg/ml. 5-Cl-PZA slows cell growth at a concentration of 25 μg/ml, although still doesn't fully inhibit cell growth with 50 μg/ml, consistant with previous observations (Zhang and Mitchison 2003) (Figure 3.2).

Proteins identified in LC/LC/MS/MS experiments

Approximately 2,000,000 MS/MS peptide fragmentation spectra were collected and analyzed over the course of 33 LC/LC/MS/MS experiments, characterizing the proteins expressed in 33 samples drawn from time courses of *M. smegmatis* growing in three different drug (INH, EMB, 5-Cl-PZA) treatments. By integrating these data with 890,000 MS/MS spectra from 27 untreated control experiments (Wang et al. 2005), we analyzed a total of 2,550 *M. smegmatis* proteins. Proteins were identified using ProteinProphet (Nesvizhskii et al. 2003) at an estimated false-positive identification rate of approximately 5%. These identified proteins represent ~40% of the ~6,500 genes expected in the *M. smegmatis* genome (personal comm. with David Graham), whose final

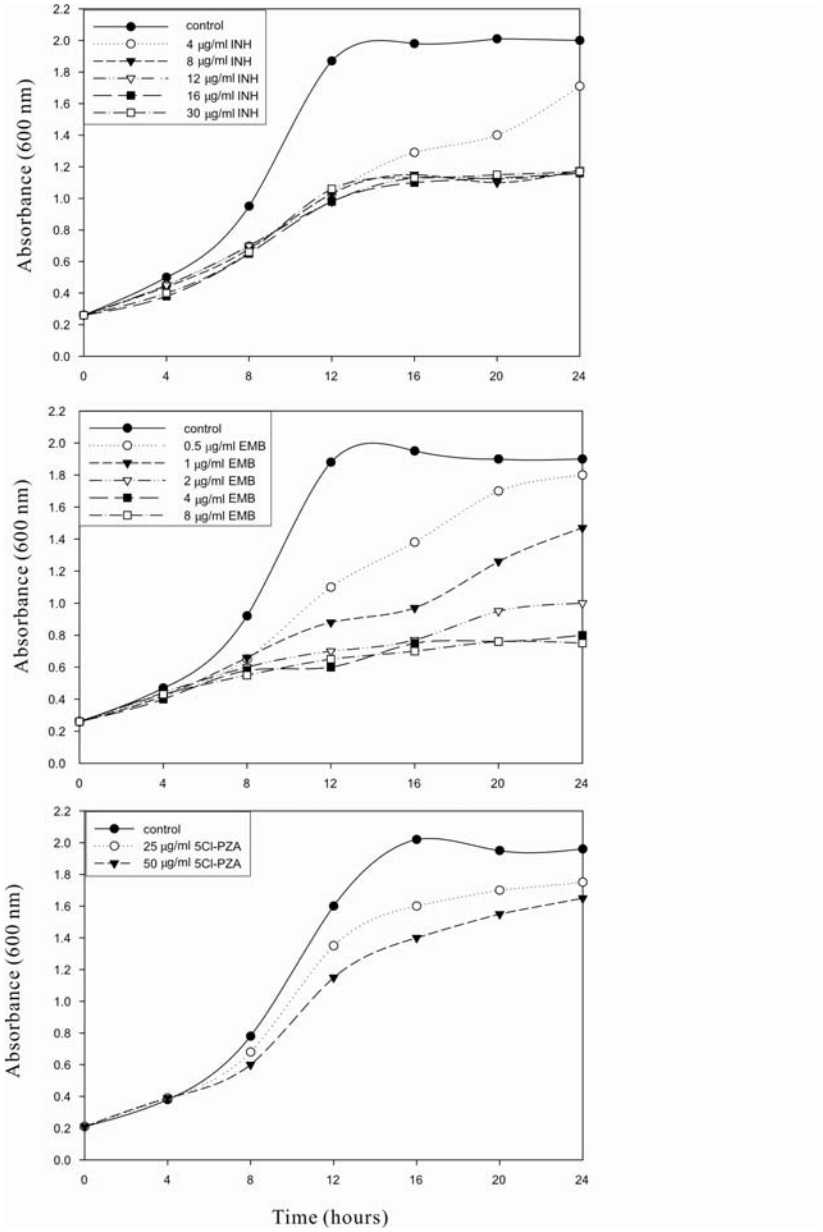


Figure 3.2 Viability of *M. smegmatis* cells following INH (A), EMB (B), and 5-Cl-PZA (C) treatments.

M. smegmatis mc^2155 cells were grown in Middlebrook 7H9 medium. INH, EMB or 5-Cl-PZA was added at the indicated concentration into cells in early-log-phase at the 0 hour time point. Cells were further incubated and monitored for growth rates once every four hours. At times indicated, small portions of the cultures were removed and the cell density was determined by OD_{600}.

sequence and annotation is currently underway at The Institute for Genomic Research (TIGR).

The global protein expression response of *M. smegmatis* to anti-TB drugs

It has been observed that the number of MS/MS spectra associated with each identified protein in shotgun proteomics experiments correlates well with protein concentrations (Liu et al. 2004). We estimated the protein abundance for each protein in drug treated *M. smegmatis* cells using this approach (Supplemental Table 3.1). Protein expression profiles derived from the fraction of sequenced peptides per protein reveal that proteins were differentially expressed in each treatment (Figure 3.3). Proteins involved in drug target pathways generally cluster together in this analysis. INH target 3-oxoacyl-[acyl-carrier-protein] synthase 2 (KasB) is in the same operon as 3-oxoacyl-[acyl-carrier-protein] synthase 1 (KasA), meromycolate extension acyl carrier protein (AcpM), malonyl CoA-acyl carrier protein transacylase (FabD), and acetyl/propionyl-CoA carboxylase (AccD6) (Figure 3.5C). Proteins encoded by genes in this operon, which are involved in the FAS-II fatty acid synthesis pathway, are coordinately up-regulated when cells are treated with INH. Fatty acid synthetase (Fas), propionyl-coA carboxylase (AccD5) and acetyl-/propionyl-coenzyme A carboxylase (AccA3), proteins in the mycolic acid synthesis pathway, are also coexpressed in this cluster as might be expected from the general perturbation to mycolic acid synthesis. In EMB treatments, the inosine-5'-monophosphate dehydrogenases GuaB1, GuaB2, and GuaB3 are up-regulated. Glutamine synthetase GlnA and GlnB are up-regulated in 5-Cl-PZA treatment, which might be explained by 5-Cl-PZA-induced amide production activating the L-glutamine catabolism pathway (Bugrim et al. 2004).

Operon-encoded proteins change expression levels coordinately

With these relative expression values, we analyzed the expression of proteins encoded in the same operons. First, as a simple check that these data are informative for measuring expression levels, a comparison of the Pearson's correlation coefficient between protein expression vectors versus the physical distance between the corresponding genes on the *M. smegmatis* genome reveals that the smaller the distance between genes oriented in the same direction, the higher the expression correlation (Figure 3.4), as expected for coordinately regulated genes.

Second, we see reasonably high coherence in the protein expression patterns of proteins in the same operon (Figure 3.5). For example, the expression levels of the proteins AtpF, AtpH, AtpA, AtpG, AtpD, and AtpC, subunits of the F0F1-type ATP synthase, are consistent in the three drug treatments and untreated control (Figure 3.5A). Figure 3.5B shows that the proteins in the RpsJ and RpsS operons, components of the 11 gene S10 ribosomal protein operon, are strongly down-regulated in all three anti-TB drug treatments in a coordinate fashion. This result may indicate that these three drugs actively inhibit protein translation in cells, or, more likely, may be a secondary effect of the cells dying.

We searched for operons that were up- or down-regulated in specific drug treatments. The proteins encoded by the KasB operon are strongly up-regulated in INH treatment (Figure 3.5C). KasB, the target of INH (Banerjee *et al.* 1994) is in the same operon as KasA, AcpM, FabD, and AccD6, which are grouped into one co-expression cluster, except AcpM (Figure 3.3B). The KasB operon was significantly up-regulated in

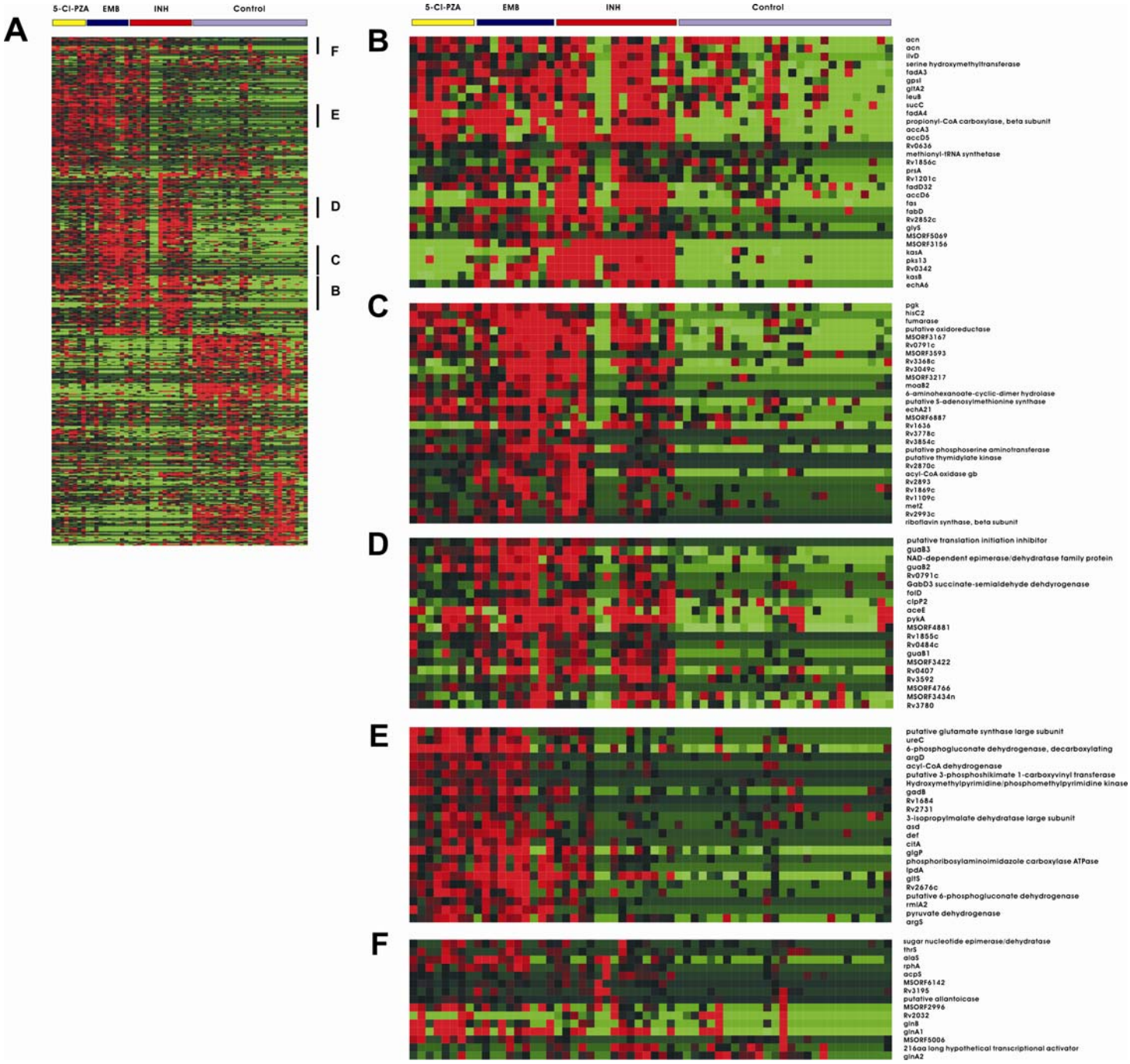

A
5-Cl-PZA
EMB
INH
Control
F
E
D
C
B
B
5-Cl-PZA
EMB
INH
Control
acn
acn
ilvD
serine hydroxymethyltransferase
fadA3
gpsI
gltA2
leuB
sucC
fadA4
propionyl-CoA carboxylase, beta subunit
accA3
accD5
Rv0636
methionyl-tRNA synthetase
Rv1856c
prsA
Rv1201c
fadD32
accD6
fas
fabD
Rv2852c
glyS
MSORF5069
MSORF3156
kasA
pks13
Rv0342
kasB
echA6
C
pgk
hisC2
fumarase
putative oxidoreductase
MSORF3167
Rv0791c
MSORF3593
Rv3368c
Rv3049c
MSORF3217
moaB2
6-aminohexanoate-cyclic-dimer hydrolase
putative S-adenosylmethionine synthase
echA21
MSORF6887
Rv1636
Rv3778c
Rv3854c
putative phosphoserine aminotransferase
putative thymidylate kinase
Rv2870c
acyl-CoA oxidase gb
Rv2893
Rv1869c
Rv1109c
metZ
Rv2993c
riboflavin synthase, beta subunit
D
putative translation initiation inhibitor
guaB3
NAD-dependent epimerase/dehydratase family protein
guaB2
Rv0791c
GabD3 succinate-semialdehyde dehdyrogenase
folD
clpP2
aceE
pykA
MSORF4881
Rv1855c
Rv0484c
guaB1
MSORF3422
Rv0407
Rv3592
MSORF4766
MSORF3434n
Rv3780
E
putative glutamate synthase large subunit
ureC
6-phosphogluconate dehydrogenase, decarboxylating
argD
acyl-CoA dehydrogenase
putative 3-phosphoshikimate 1-carboxyvinyl transferase
Hydroxymethylpyrimidine/phosphomethylpyrimidine kinase
gadB
Rv1684
Rv2731
3-isopropylmalate dehydratase large subunit
asd
def
citA
glgP
phosphoribosylaminoimidazole carboxylase ATPase
lpdA
gltS
Rv2676c
putative 6-phosphogluconate dehydrogenase
rmlA2
pyruvate dehydrogenase
argS
F
sugar nucleotide epimerase/dehydratase
thrS
alaS
rphA
acpS
MSORF6142
Rv3195
putative allantoicase
MSORF2996
Rv2032
glnB
glnA1
MSORF5006
216aa long hypothetical transcriptional activator
glnA2

Figure 3.3

Figure 3.3 Hierarchical clustering of *M. smegmatis* proteins identified across 60 shotgun proteomics experiments.

The data set is the fraction of sequenced peptides for proteins identified more than 20 times across 60 shotgun experiments. (A) Proteins were clustered into groups in different treatments, black bars on the right indicate induced protein expression in drug treatments. (B-F) Proteins with expression levels up-regulated in anti-TB drug treatments are involved in the drug-target pathways.

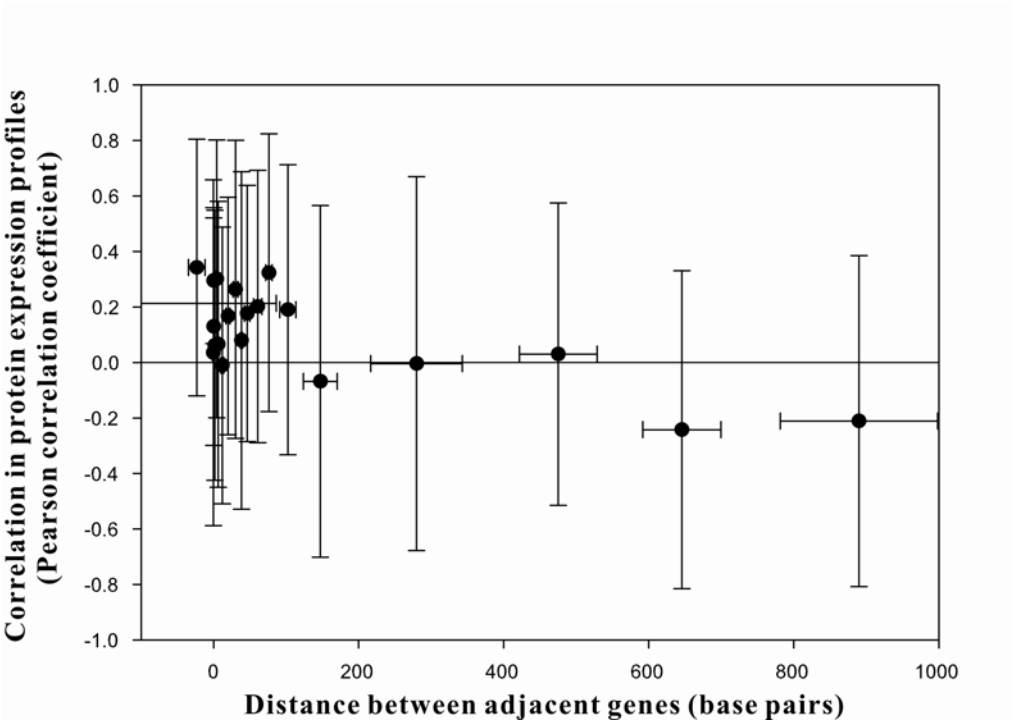


Figure 3.4 Pearson's correlation coefficient between identified proteins correlate with physical distance between genes oriented the same direction.

Pearson's correlation coefficient is calculated based on the fraction of identified peptides per protein across all 60 shotgun experiments. The distance is the number of base pairs between adjacent genes transcribed in the same direction.

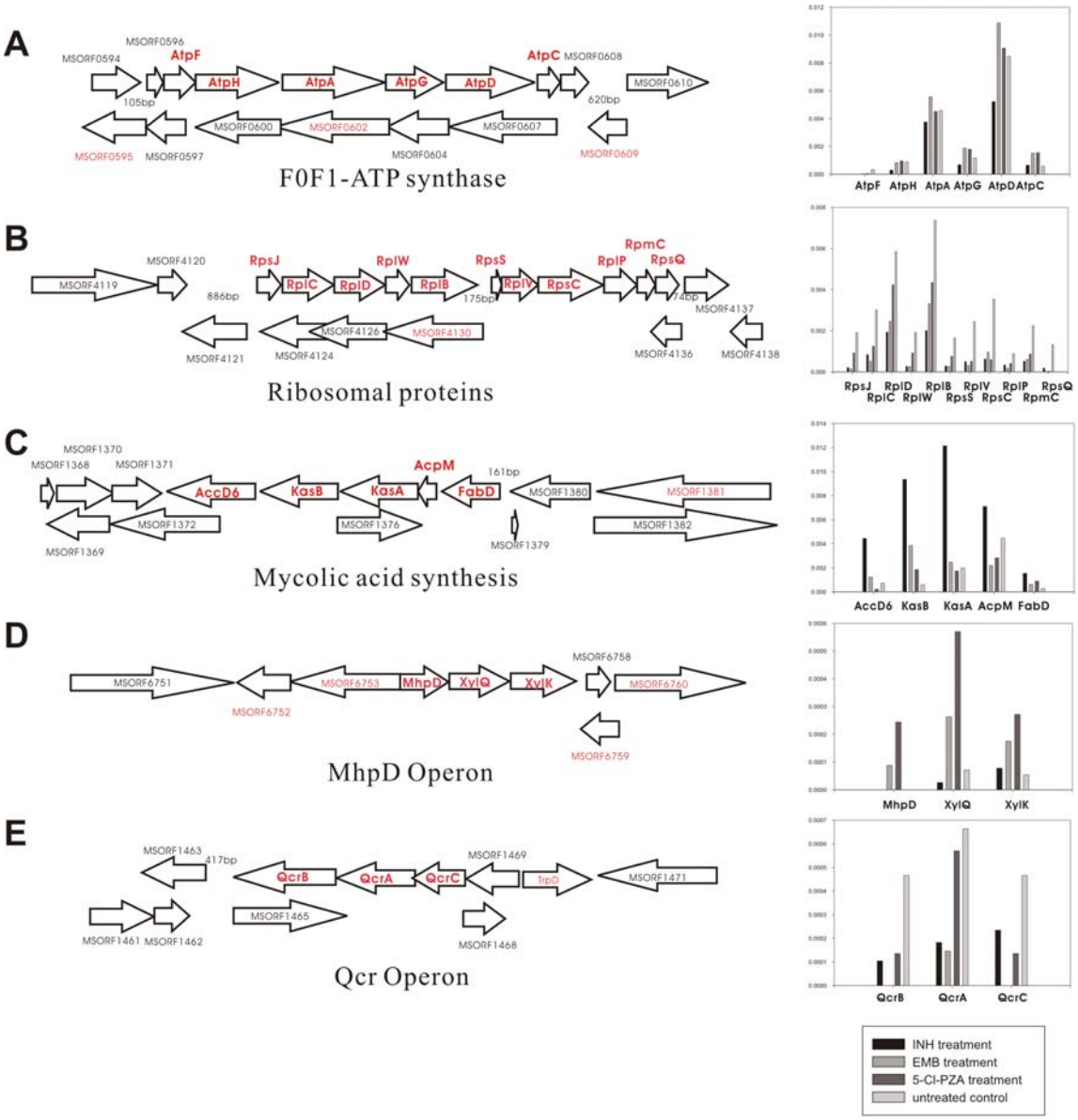


Figure 3.5 Operon-encoded protein expression level changes coordinately in drug treatments.

Proteins with red colors were identified in our experiments, and red bold proteins are in the same operon. Proteins in the ATPase operon (A) express consistently in all treatments; proteins in RpsJ, RpsS operons (B) are down-regulated in three drug treatments; the KasB operon (C), and the MhpD operon (D) are up-regulated in INH, 5-Cl-PZA treatments; proteins in QcrABC operon (E) are down-regulated in EMB treatment. The y axes in column charts (right) represent relative protein abundance.

quantitative analysis of INH-induced proteomic changes (Hughes et al. 2006). These proteins are known to be in the mycolic acid synthesis pathway (Wilson *et al.* 1999), the main component of the cell wall in mycobacteria. We reason that the INH-induced up-regulation of these enzymes in FASII system suggests the bacteria is up-regulating mycolic acid synthesis enzymes to counteract the inhibitory effects of INH. Similarly, 4 hydroxy-2-oxovalerate aldolase (XylQ), acetyldehyde dehydrogenase (XylK), and 2-keto-4-pentenoate hydratase (MhpD) are up-regulated in 5-Cl-PZA treatment (Figure 3.5D). This operon is involved in secondary metabolite biosynthesis, transport, and catabolism (http://www.ornl.gov/); the operon's specific induction in 5-Cl-PZA-treated cells may implicate this system or those functionally linked to it as PZA targets. The ubiquinol-cytochrome C oxidoreductase QcrABC operon is down-regulated in EMB treatment (Figure 3.5E); this protein complex is an integral membrane enzyme that catalyzes electron transfer from a quinol to a c-type cytochrome (Yu et al. 1995).

Protein differential expression in *M. smegmatis* in anti-TB drug treatments

To characterize the cellular response to anti-TB drugs, we measured differential expression of identified proteins by counting peptides identified from each protein. The significance of differential expression was calculated as a Z score from the difference in peptide coverage for a given protein between two experiments (Lu et al. 2005b). The Z scores of identified proteins across 32 experiments were calculated (Supplemental Table 3.2), and proteins showing significant expression changes were chosen at the 99% confidence level ($|Z| \geq 2.58$). The differential protein profiles provide information on which proteins are responsive in drug treatments and may belong to drug target pathways.

We evaluated this strategy by calculating differential protein expression for INH-treated cells, then testing how the known targets in the mycolate biosynthesis pathway behaved. Figure 3.6 shows that mycolic acid biosynthetic proteins are effectively identified by this strategy. While most cellular proteins are in the range of $|Z|<2$, the 22 mycolic acid biosynthetic proteins are predominantly with $|Z|>2$, with the majority up-regulated.

Using this strategy, we tested for proteins strongly up- or down-regulated in each drug treatment. Each significantly differentially expressed protein was associated with a functional category and the enrichment for each functional category was calculated. The enriched categories provide biological interpretation of the mass spectrometry-based measurement of differential expression. Functional enrichment for each drug treatment is shown in Figure 3.7. Translation, energy production, and protein export are all down-regulated in the three drug treatments. These three drugs each inhibit cell growth (Figure 3.7B), apparent in the interruption of basic cellular processes. INH up-regulates lipid transport and metabolism 1.2-fold (Figure 3.7A), as expected for cells whose fatty acid synthesis (FASII) and mycolic acid synthesis are inhibited by INH. Lipid transport and metabolism is also up-regulated in EMB and 5-Cl-PZA treatments, in accordance with previous observations (Takayama and Kilburn 1989; Zhang 2005; Zhang and Mitchison 2003). In addition, amino acid metabolism and transport is up-regulated in the three drug treatments. Some functional categories are under-represented, such as transcription and cell wall biogenesis, possibly stemming not from drug action but from the incomplete sampling of low-abundance and membrane proteins by shotgun proteomics.

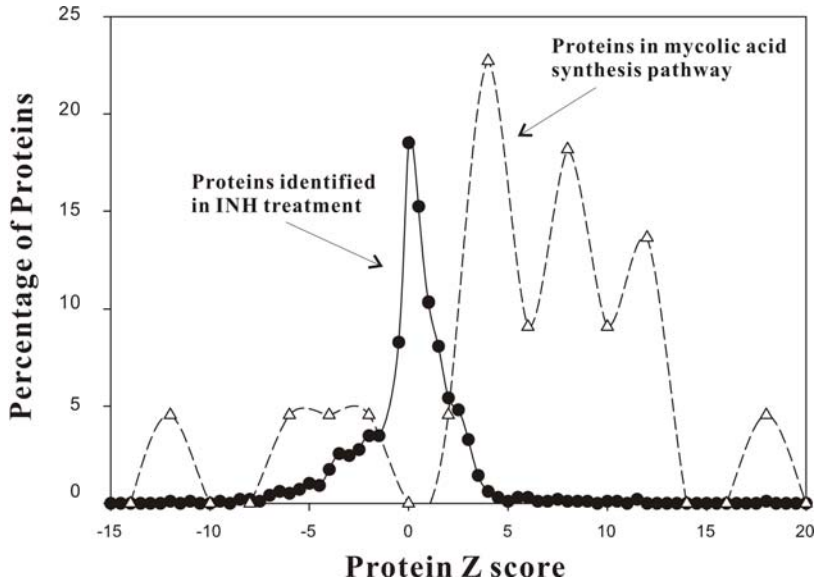


Figure 3.6 Proteins in the INH target pathway tend to have higher |*Z*| scores in INH treatment.

Solid circles represent proteins shown in INH treatment, empty triangles represent proteins in the pathway of mycolic acid synthesis.

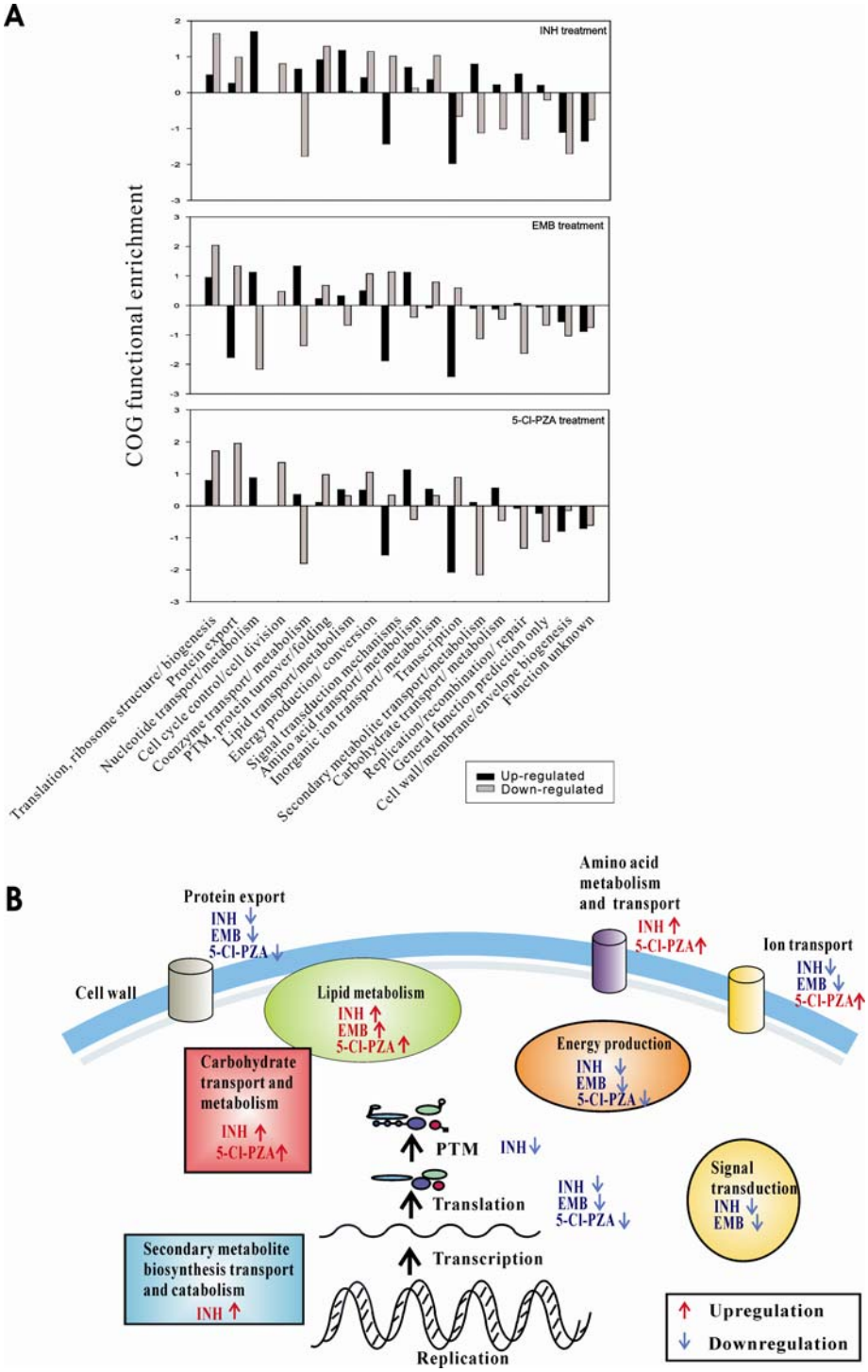

A
INH treatment
EMB treatment
5-Cl-PZA treatment
COG functional enrichment
Translation, ribosome structure/ biogenesis
Protein export
Nucleotide transport/metabolism
Cell cycle control/cell division
Coenzyme transport/ metabolism
PTM, protein turnover/folding
Lipid transport/metabolism
Energy production/ conversion
Signal transduction mechanisms
Amino acid transport/ metabolism
Inorganic ion transport/ metabolism
Secondary metabolite transport/metabolism
Carbohydrate transport/ metabolism
Transcription
Replication/recombination/ repair
General function prediction only
Cell wall/membrane/envelope biogenesis
Function unknown
Up-regulated
Down-regulated
B
Protein export
INH
EMB
5-Cl-PZA
Amino acid metabolism and transport
INH
5-Cl-PZA
Ion transport
INH
EMB
5-Cl-PZA
Cell wall
Lipid metabolism
INH
EMB
5-Cl-PZA
Energy production
INH
EMB
5-Cl-PZA
Carbohydrate transport and metabolism
INH
5-Cl-PZA
PTM
INH
Translation
INH
EMB
5-Cl-PZA
Signal transduction
INH
EMB
Transcription
Secondary metabolite biosynthesis transport and catabolism
INH
Replication
Upregulation
Downregulation

Figure 3.7

Figure 3.7 Protein expression levels change in drug treatments.

Proteins were selected as significantly differential expressed with 99% confidence. (A) COG functions for the proteins differentially expressed in each drug treatment. Black bars represent up-regulated proteins ($Z \geq 2.58$) in each drug treatment; grey bars represent down-regulated proteins ($Z \leq -2.58$). For the proteins in each COG category, we plot the relative difference in expression level, calculated as log ($N_{up,\,COGx}/N_{up}/N_{COGx}/N$), where $N_{up,\,COGx,}$ and N_{up} represent the number of up-regulated proteins with COG*x* function and up-regulated proteins in each drug treatment, N_{COGx} and N are the number of observed proteins with COG*x* function and the number of observed proteins for each drug treatment. PTM, post-translational modification. (B) The significant functional enrichments in hypergeometric distribution are shown in the plot. Proteins with red color represent the protein expression levels were up-regulated in drug treatments, proteins with blue color represent down-regulation.

Figure 3.8 presents a visual summary of the drug-induced differential expression of proteins. Across 32 experiments, we identified 485 proteins differentially expressed with 99% confidence ($|Z| \geq 2.58$). For each protein, a point is plotted within the triangle at a position signifying the relative enrichment of the protein across the treatments (Tringe et al. 2005). The relative peptide fractions in the three drug treatments were normalized to add up to 1 in order to indicate the relative expression, each fraction calculated as $f_{drug,i} = \frac{Fraction_{drug,i}}{\sum_{all\_drugs,i} Fraction_{j,i}}$, where $Fraction_{drug,i}$ represents the fraction of identified peptides for the protein i in total peptides observed in the drug's shotgun experiments. Proteins plotted in the middle of the triangle are equally abundant in all three drug treatments. Proteins sitting in one of the corners are highly induced in one drug treatment, and proteins that appear along one of the edges are induced in only two of the three drug treatments.

The effect of INH

Using Figure 3.8, we first summarize the proteins specifically induced in INH-treated cells. Proteins plotted in the INH-specific corner are involved in the biosynthesis of mycolic acid (Mawuenyega et al. 2005) (Raman et al. 2005). These proteins include the INH targets KasB and enoyl-ACP-reductase (InhA) (Sinha et al. 2005). Alkylhydroperoxidase (AhpD, an element of the peroxiredoxin defense against oxidative stress), mycolic acid synthase (UmaA1), and acetyl-coA acyltreansferase (FadA2, involved in lipid degradation) are in the corner as up-regulated. In addition, proteins in FAS-II system, such as, KasA, AcpM, FabD, AccD6, polyketide synthase (Pks13), fatty-acid-CoA ligase (FadD32), and 3-oxoacyl-[acyl-carrier-protein] reductase (FabG4), are up-regulated as expected in INH treatments. The increased protein production results in

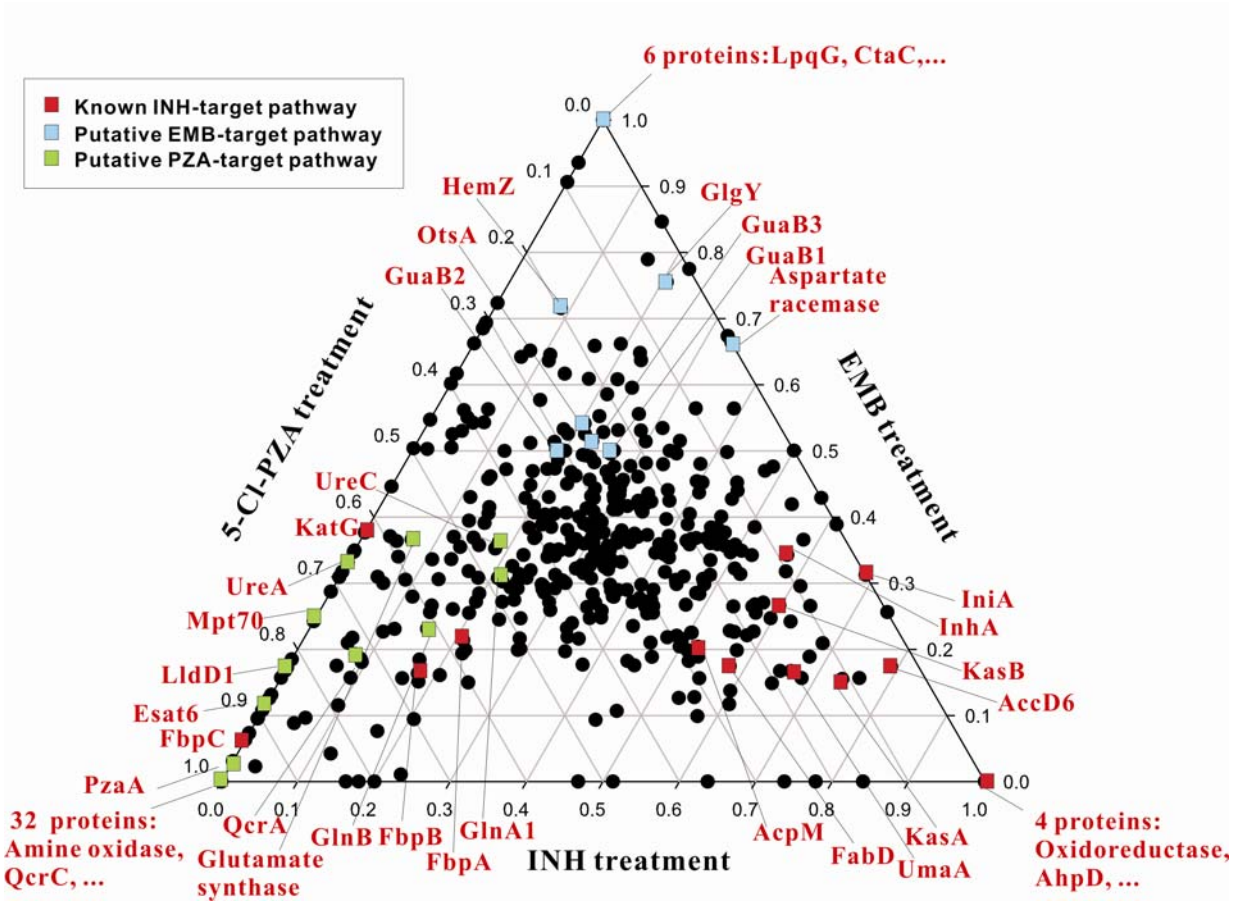


Figure 3.8 Specific protein expression enrichments in drug treatments.

Three-way comparison of INH, EMB, and 5-Cl-PZA treatment in terms of protein relative abundance in *M. smegmatis*. Each dot represents relative abundance of a protein in three drug treatments so that proximity to a vertex is proportional to the level of enrichment in the respective sample. Colored squares are annotated proteins: red squares represent known proteins in INH pathway (mycolic acid synthesis pathway); blue squares represent proteins discussed in the results of the EMB experiments; green squares are proteins discussed in the results of the 5-Cl-PZA treatment.

the accumulated fatty acid pathway precursors. Fatty acid synthetase (Fas), which catalyzes the formation of long-chain fatty acids, is up-regulated. The isoniazid inducible protein (IniA) (Colangeli et al. 2005), essential for activity of an efflux pump that confers drug tolerance to both INH and EMB, is also up-regulated. INH up-regulates the enzymes in FASII system, as reflected in the proteomics data. It suggests that the induction of these proteins is the consequence of a regulatory feedback mechanism that bacteria up-regulate mycolic acid synthesis enzymes to conteract the inhibitory effects of INH.

Peroxidase/catalase (KatG), which activates INH, is strongly down-regulated. The end products mycolyltransferase (FbpA, B, C) of mycolic acid pathway, the three subunits of antigen85 / fibronectin-binding protein, are down-regulated, consistent with previous observations (Cabusora et al. 2005; Wilson et al. 1999). As a consequence of INH activity, mature mycolates are not produced and become progressively depleted.

The effect of EMB

Next, we summarize specific responses to EMB. EMB targets the mycobacterial cell wall. The target genes are arabinosyl transferases (EmbABC), which act on the polymerization of arabinose into the arabinan of arabinogalactan (AG) and lipoarabinomannan (Jungblut et al. 1999), the major polysaccharide of the mycobacterial cell wall (Deng et al. 1995; Mikusova et al. 1995; Takayama and Kilburn 1989). In our experiments, proteins involved in FASII system are up-regulated in EMB treatments such as KasB, Pks13, and Fas. In contrast, FbpA and FbpC are down-regulated, consistent with EMB indirectly inhibiting mycolic acid synthesis by limiting the availability of arabinan for the mycolic acids to attach to (David et al. 1989), and triggering a cascade of changes in the lipid metabolism of mycobacteria.

In Figure 3.8, proteins shown in the corner of EMB include 7 membrane proteins, such as penicillin-binding proteins (essential membrane-bound cell-wall synthesizing enzyme), aspartate racemase (cell wall biogenesis), transmembrane alanine and glycine rich protein (peptidoglycan biosynthesis), lipoprotein (lpqG), and transmembrane cytocheome C oxidase CtaC (aerobic respiration). In addition, maltooligosyltrehalose synthase (GlgY, involved in trehalose biosynthesis), enoyl-CoA hydratase (beta oxidation of fatty acids), and cystathionine/methionine gamma-synthase/lyase. Ferrochelatase (HemZ) is in the operon with two genes (mabA and inhA) involved in mycolic acid biosynthesis. Glutamyl-tRNA synthetase (GltS) and IMP dehydrogenase (GuaB1, Gua2, Gua3, involved in GMP biosynthesis) are up-regulated, as glutamate and cAMP are necessary for galactose to be incorporated into cell wall (Raychaudhuri et al. 1998). Glutamate-1-semialdehyde 2,1-aminonutase (HemL) is up-regulated can be explained as EMB can help the glutamate-efflux (Radmacher et al. 2005). Zinc metalloprotease, a lethal factor, is capable of cleaving members of the membrane-associated protein (MAP) kinase family in macrophages (Vitale et al. 1998). Monooxygenase (MSORF3583 and MSORF 2318) and oxidoreductase (lpdA and MSORF8662) are up-regulated due to the stress of the EMB treatment. The up-regulation of pyruvate carboxylase (Pca, an anaplerotic enzyme) might be because the anaplerotic reactions play important roles in the biochemical differentiation of mycobacteria into non-replicating stages (Mukhopadhyay and Purwantini 2000).

The effect of PZA

Finally, we summarize the specific responses to 5-Cl-PZA. PZA is a prodrug, which has to be activated in the cell. The active derivative of PZA is pyrazinoic acid,

which is preferentially accumulated in an acidic pH. Mutations in the gene encoding pyrazinamidase/nicotinamidase (PncA) cause resistance to PZA (Scorpio and Zhang 1996). The accumulation of pyrazinoic acid results in the death of cells. Relatively little is known about the targets of PZA, although PZA inhibits L-tryptophan catabolism (Saito et al. 2000) and increases catabolism of L-glutamine (Bugrim et al. 2004). PZA blocks energy transport by disrupting membrane functions (Zhang 2005; Zhang and Mitchison 2003). PZA has been shown to inhibit fatty acid synthesis (Boshoff et al. 2002; Zimhony et al. 2000) and PZA elevates the NAD^+ levels in cells (Shibata et al. 2001).

The proteins shown in the corner of 5-Cl-PZA include PzaA, a homolog of PncA, which is up-regulated, possibly indicating an activity related to PzaA's function in activating PZA by hydrolyzing PZA to pyrazinic acid. Glutamine synthetase GlnA1 and GlnB, playing a central role in nitrogen metabolism and the formation of a poly-L-glutamate/glutamine cell wall structure (Harth et al. 2000), are up-regulated in 5-Cl-PZA treatment, which are responsible for NH_2^+ and energy status (Bugrim et al. 2004). The up-regulation of urea amidohydrolase (UreA and UreC) can be explained as these widely distributed proteins from microorganisms to vertebrates produce hydrogen peroxide plus aldehyde when catabolizing endogenous or xenobiotic amines (Carpene et al. 2005). In addition, MSORF2805 (Rv1626 ortholog, nitrogen regulation protein), cyanate lyase LlD1, formamidase, isocitrate dehydrogenase (Icd2), and Rv0462 ortholog (MSORF1225, involved in energy production) are up-regulated. Those proteins are involved in nitrogen regulation and energy metabolism. Amine oxidase, amino acid transport system protein, aminotransferase, and oxidoreductase are up-regulated.

To further investigate PZA effects, we plotted the expression changes in the GO functional categories for the significantly differential expressed proteins ($|Z| \geq 2.58$) in the 5-Cl-PZA drug treatments (Figure 3.9). We expect the most enriched categories might correspond to the pathways of PZA drug targets. Among up-regulated proteins, carboxylic acid metabolism, organic acid metabolism, nitrogen compound metabolism, and amino acid metabolism are significantly over represented. In contrast, the GO pathways of protein biosynthesis and protein metabolism are significantly enriched among down-regulated proteins. It would appear from the proteomics data that PZA's primary specific effects are to perturb cellular nitrogen metabolism, carboxylic acid, and organic acid metabolism.

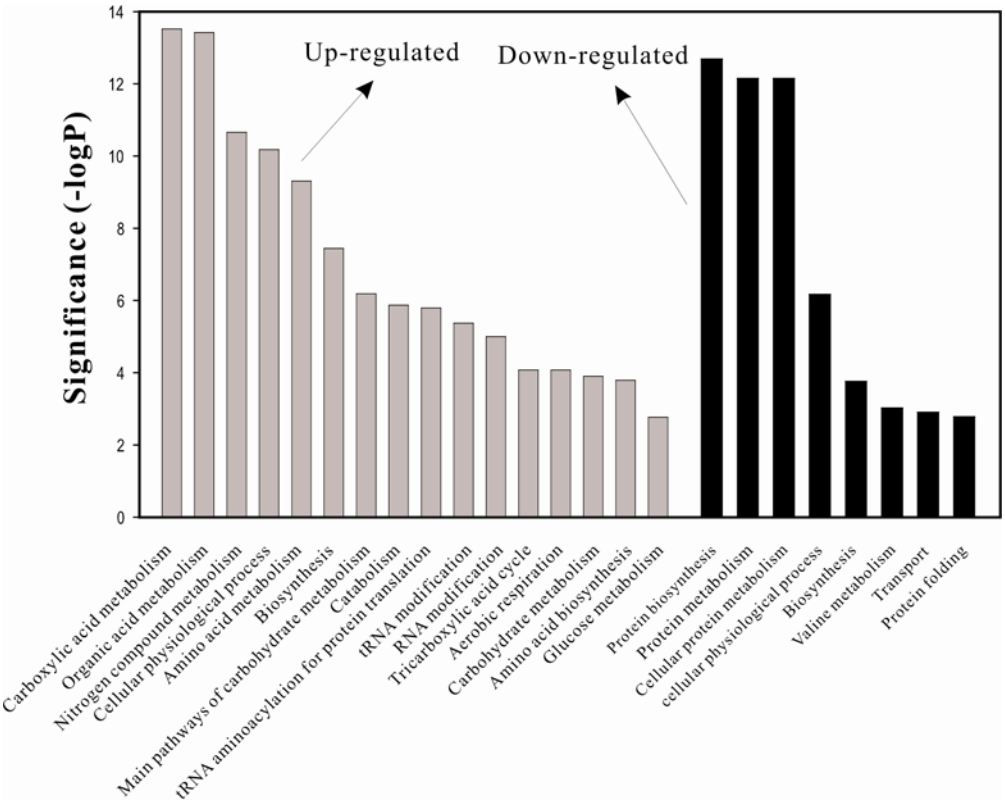


Figure 3.9 The significant GO functional categories in proteins identified in 5-Cl-PZA treatments.

Hypergeometric distribution was calculated on the proteins show significantly differential expression ($|Z| \geq 2.58$) in 5-Cl-PZA treatments. The y axis represents the log-scale of the P-values in hypergeometric distribution. Gray bars represent the GO functional categories for the significant up-regulated proteins; black bars represent the GO functional categories for the significant down-regulated proteins.

DISCUSSION

The rapid progress in proteomics technology makes large-scale study of 100's to 1000's of proteins possible. In the present analysis, we used this technology to shed light on drug action. An important factor in developing a new drug is the criterion that the compound must provide selective inhibition of the intended target. Protein profiling is useful for this confirmation through comparison of protein expression profiles obtained in response to novel inhibitors with signature protein expression profiles of drugs of known modes of action.

The shotgun proteomics data generated in the process also provides an efficient method for annotating the expressed proteome. In all, we provide experimental support for 2550 proteins (Figure 3.10). Their COG categories (Figure 3.10A) indicate significant enrichment for proteins of energy production (6.5%), amino acid transport and metabolism (6%), lipid transport and metabolism (4.8%), as well as a large fraction of uncharacterized proteins (44%). There is a clear enrichment over random expectation (Table 3.1) in proteins of protein biosynthesis, amino acid biosynthesis and metabolism, nucleotide biosynthesis and metabolism, and carbohydrate metabolism. These set of proteins show depletion for DNA recombination and RNA transcription. Thus, additional proteomics evidence for these proteins could be generated by targeting transcription regulation proteins.

For each protein identified, we calculated the number of identified peptides and examined the distribution of these values across the set of 2550 proteins (Figure 3.11). While the majority of proteins were observed a small number of times, an appreciable fraction of the proteins (35%) had 10 or more repeat observations of associated peptides,

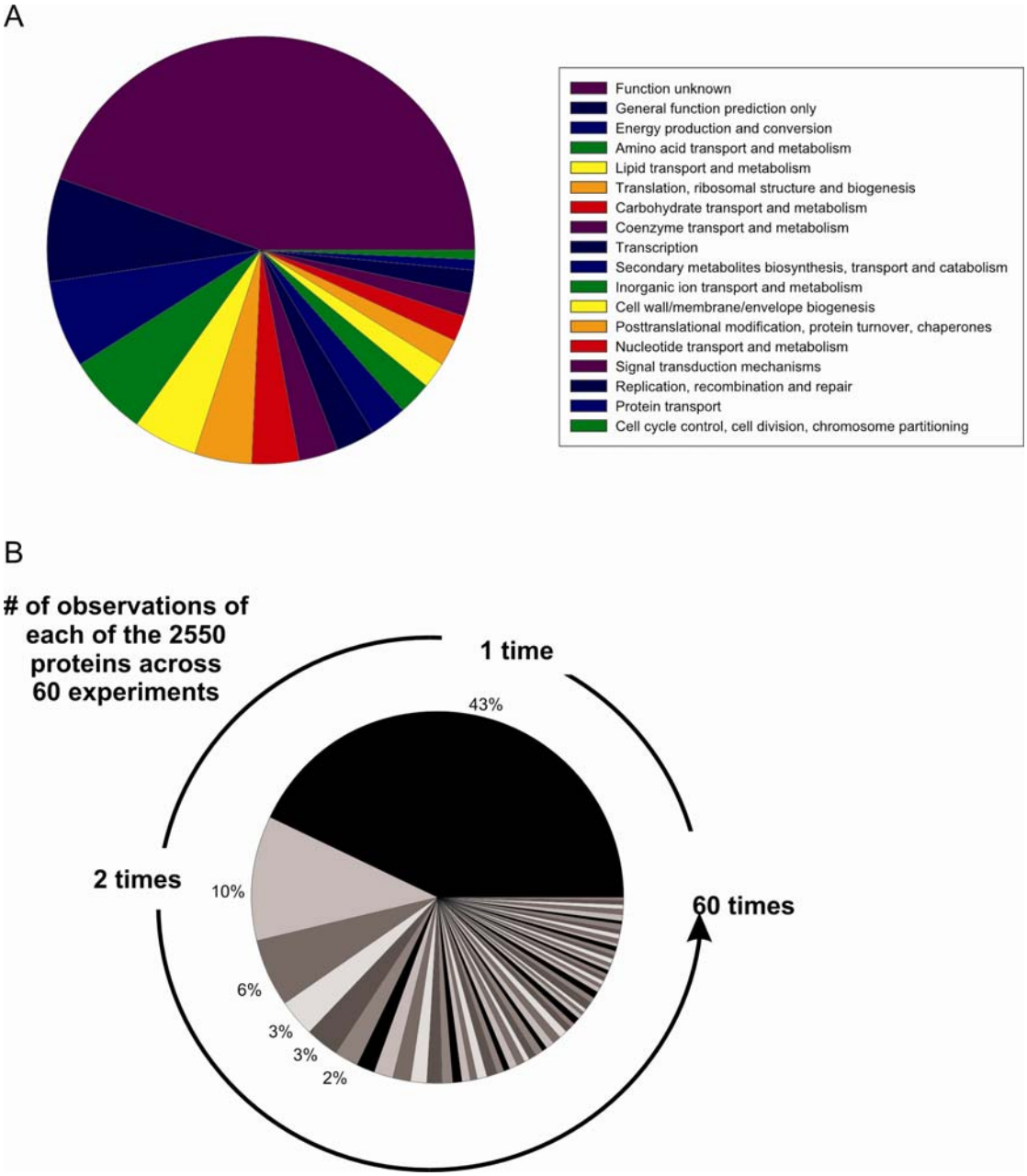


Figure 3.10 The distribution of observations for the 2550 proteins across all experiments and the associated protein functions for the complete set of proteins.

(A) The distribution of observation of each of the 2550 proteins across identified across all 60 shotgun experiments. (B) The COG functional distribution for the 2550 identified proteins.

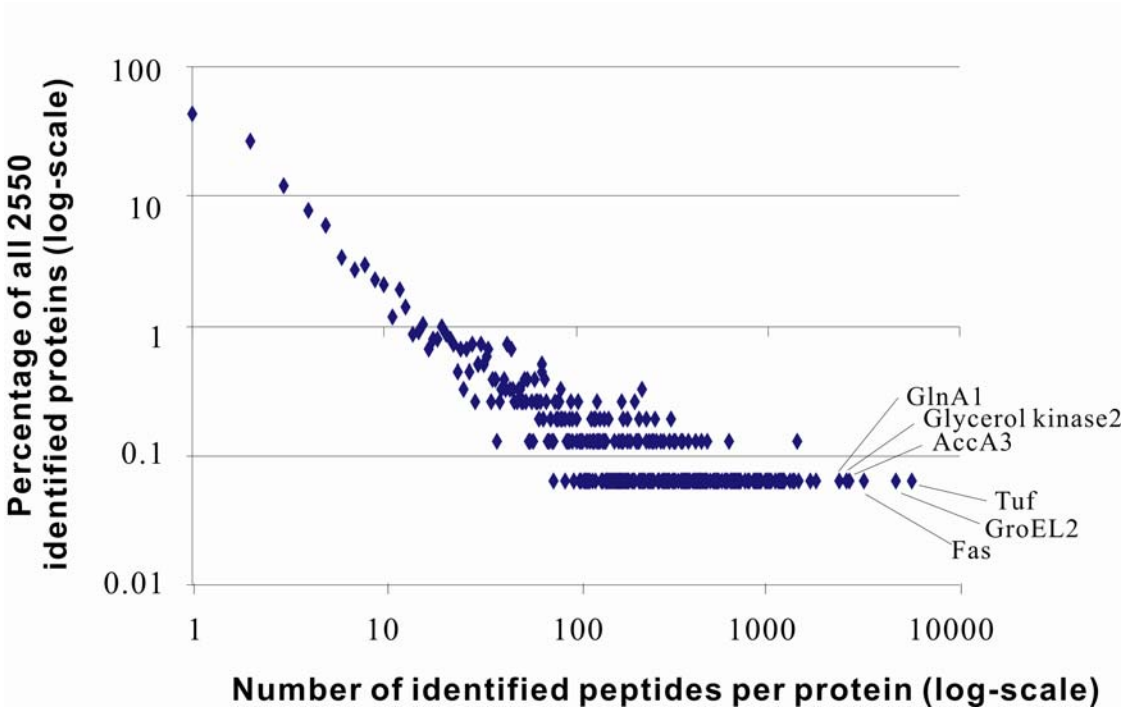


Figure 3.11 The total number of identified proteins is anti-correlated with the number of observed peptides.

The x axis represents the number of all peptides sequenced for protein; the y axis represents the percent of the 2550 identified proteins.

GO_ID	GO_annotation	*p*(fraction)	F(significance)	Functional Enrichment
GO:0009058	biosynthesis	0.306084819	9.00E-13	1.399541409
GO:0006082	organic acid metabolism	0.142593731	2.53E-08	1.516309413
GO:0019752	carboxylic acid metabolism	0.141979103	3.22E-08	1.513355563
GO:0006520	amino acid metabolism	0.088506454	3.97E-08	1.679523273
GO:0006807	nitrogen compound metabolism	0.09956976	8.66E-08	1.615056723
GO:0006412	protein biosynthesis	0.063921328	9.28E-08	1.796972453
GO:0008652	amino acid biosynthesis	0.057775046	8.26E-06	1.684071305
GO:0044267	cellular protein metabolism	0.145052243	1.96E-05	1.378813559
GO:0019538	protein metabolism	0.145666872	2.42E-05	1.372995781
GO:0009117	nucleotide metabolism	0.045482483	0.000184143	1.634130752
GO:0005975	carbohydrate metabolism	0.072526122	0.000276921	1.471976638
GO:0006164	purine nucleotide biosynthesis	0.015980332	0.000497806	2.02952183
GO:0006092	main pathways of carbohydrate metabolism	0.025814382	0.000585167	1.77985843
GO:0006163	purine nucleotide metabolism	0.017209588	0.000643284	1.963079151
GO:0006418	tRNA aminoacylation for protein translation	0.012907191	0.000841035	2.093951094
GO:0009165	nucleotide biosynthesis	0.035648433	0.000843502	1.630032619
GO:0009152	purine ribonucleotide biosynthesis	0.012292563	0.002548035	1.978783784
GO:0006313	DNA transposition	0.019053473	0.999924318	0.1418483
GO:0006310	DNA recombination	0.038106945	0.999998691	0.2836966
GO:0006355	regulation of transcription, DNA-dependent	0.132145052	1	0.541992458
GO:0045449	regulation of transcription	0.133988937	1	0.54461939
GO:0006351	transcription, DNA-dependent	0.134603565	1	0.572251018
GO:0006350	transcription	0.136447449	1	0.574421719

Table 3.1 The significant enrichment of GO functional categories for proteins differentially expressed in drug treatments.

The significance is calculated using a two-tailed binomial distribution, for rare GO functional categories with p <0.05, the significance is calculated based on Poisson statistics. The enrichment was calculated as $f = \frac{n}{N * p}$, where n represents the total number of the identified proteins, k is the number of identified proteins with the functional category j, p is the fraction of proteins in the functional category j for the whole genome. The GO functional categories with significance F < (1/350) represent enrichment for proteins expressed differentially in drug treatments; F > (1-1/350) represent depletion.

with the highest sampling observed for the protein Tuf, whose peptides we observed a remarkable 5603 times; other highly observed protein include GroEL2, Fas, AccA3, glycerol kinase 2 and GlnA1. As peptide counts can be used to estimate protein abundances (Lu et al. 2005b), these data indicate that, in principle, expression differences of up to 4 orders of magnitude might be measured by this approach.

Proteomics includes not only protein identification but also protein quantification. Quantitative proteomics is capable of measuring protein expression levels and the differential expression of proteins. The quantitative proteomics provides us the valuable information on action mode of anti-TB drugs. In order to explore alterations in *M. smegmatis* protein profiles, we have, by using LC/LC/MS/MS, generated signature profiles of *M. smegmatis* in response to treatment with INH, EMB, and 5-Cl-PZA. The protein profiles of *M. smegmatis* during exposure to drugs are direct consequences of drugs. We expect the protein expression changes to predominantly reflect differential expression of metabolite biosynthetic enzymes. We have compared and contrasted the response profiles to the three drugs and distinguish between these treatments, thereby providing insight into the differences in the mechanisms of action of these three drugs. Drugs selectively induce changes in the expression of genes coding for enzymes that comprise the affected pathway, especially before a more generalized stress response ensues.

We can reveal the full picture of each component in cells changes under different conditions. In INH treatment, proteins involved in mycolic acid synthesis pathway are up-regulated, however, some proteins are down-regulated, such as transporter proteins

FbpABC. We reason that the INH-induced up-regulation of enzymes in FASII system suggests the bacterium is up-regulateing mycolic synthesis enzymes to counteract the inhibition effects of INH. Induced proteins could be predicted to either compensate for inhibition of the target pathway or respond to the toxic effect of the drug, and we expect that the protein profiles can serve as a fingerprint of a given drug's mode of action. In quantitative proteomic analysis of the INH-induced changes in *Mycobacteria bovis* (Hughes et al. 2006), the KasB operon, and bacterioferritin (BfrB), Antigen 84 (Wag31) are significantly up-regulated, while, D-alanine:D-alanine ligase (Ddla, involved in peptidoglycan biosynthesis) and acyl-ACP desaturase (Des2) are markedly down-regulated. The resulting protein expression profiles would not only serve as a kind of signature of the drug used, but would, in the case of inhibitors whose modes of action are unknown, incriminate the affected pathways and perhaps the specificly targeted enzymes within the pathway, in EMB and 5-Cl-PZA treatments.

MATERIAL AND METHODS

Growth and drug treatment of *M. smegmatis* cells

M. smegmatis mc^2155 bacteria were grown with agitation at 37°C in Middlebrook7H9 broth (Difco, supplemented with 10% Middlebrook OADC enrichment , 0.2% glycerol and 0.05% Tween-80) to mid-exponential phase (Jacobs et al. 1991). Cultures for experimental treatment were initiated by diluting 1:200 into fresh 7H9 media and grown to early log phase ($OD_{600} \approx 0.3$) with shaking in 5% CO2 atmosphere at 37°C. Drug treatments were begun by adding filtered stock solutions of INH (10 mg/ml, sigma) or EMB (10 mg/ml, sigma) to achieve the following final concentrations 4, 8, 12, 16, 30 μg/ml for INH and 0.5, 1, 2, 4, 8 μg/ml for EMB (Wilson et al. 1999). Minimum inhibitory concentrations (MICs) were measured using the microbroth dilution technique (Parish and Stoker 1998). 5-Cl-PZA treatment was performed in Middlebrook 7H9-based medium adjusted to pH 5.6 with HCl when cultures were grown from OD600 0.2 (Phetsuksiri B 1999; Wade and Zhang 2004), with final 5-Cl-PZA concentrations of 25 and 50 μg/ml . Samples were collected at selected intervals as in (Wilson et al. 1999). Other than 25 samples from (Wang et al. 2005), 2 more untreated control samples were collected at OD600 of 1.9. Samples were centrifuged at 12,000 g for 30 min, suspended in ice-cold lysis buffer (25mM Tris HCL pH7.5, 2.5mM DTT, 1.0mM EDTA, 0.02%(w/v) Brij35, 1X Calbiochem Protease Inhibitor Cocktail Set I (CPICSI)) (1ml/g cell pellet), and disrupted by bead-beating with 1mm glass beads (Parish and Stoker 2001; Parish 2001; Primm et al. 2000). Cell lysates were clarified by centrifugation at 20,000g for 30 min, with typical protein concentrations of 10mg/ml.

Preparation and LC/LC/MS/MS analysis of *M. smegmatis* peptides

M. smegmatis soluble protein extracts were diluted in digestion buffer (50mM Tris HCL pH8.0, 1.0M urea, 2.0mM $CaCl_2$), denatured at 95°C for 15 min, and digested with sequencing grade trypsin (Sigma) at 37°C for 20 hr. Untreated control samples were analyzed as in (Wang et al. 2005). Tryptic peptide mixtures were separated by automated two dimensional-high performance liquid chromatography. Chromatography was performed at 2 μl/min with all buffers acidified with 0.1% formic acid. Chromatography salt step fractions were eluted from a strong cation exchange column (SCX) with a continuous 5% acetonitrile (ACN) background and 10 minute salt bumps of 5, 20, 60, and 900 mM ammonium chloride. Each salt bump was eluted directly onto a reverse phase C18 column and washed free of salt. Reverse phase chromatography was run in a 125 minutes gradient from 5% to 50% ACN, then purged at 95% ACN. Peptides were analyzed online with electrospray ionization (ESI) ion trap mass spectrometry (Corbett et al.) (Link *et al.* 1999; Washburn *et al.* 2001) using a ThermoFinnigan Surveyor/DecaXP+ instrument. For drug treated samples, gas phase fractionation (GPF) was used to achieve maximum proteome coverage and increase coverage of low abundant proteins (Yi et al. 2002). Three sequential LC/LC/MS/MS analyses were performed, spectra collected from different mass/charge (*m/z*) ranges (300–650, 650–1000, and 1000–1500 *m/z*) for data-dependent precursor ion selection. For each MS spectra, the 5 tallest individual peaks, corresponding to peptides, were fragmented by collision-induced dissociation with helium gas to produce MS/MS spectra. Fragmentation data from each set of three runs were combined for peptide analysis.

Protein identification

For peptide and protein identification purposes, we used a database of 8968 potential *M. smegmatis* coding sequences described in (Wang et al. 2005). Proteins were identified from the resulting peptide MS/MS fragmentation spectra by searching against the custom *M. smegmatis* predicted protein database using the program BioWorks 3.1. The unique and total number of peptides for each identified protein in a given proteomics experiment were obtained using PeptideProphet (Keller et al. 2002) and ProteinProphet (Nesvizhskii et al. 2003), thresholding the false positive protein identification rate to less than 5%.

Protein quantification

To estimate the abundance of each protein i, we employed the APEX technique (Lu et al. 2005b). We counted the number (n_i) of identified peptides and calculated relative protein abundance f_i for that experiment as $f_i = n_i / N$, where N is the total number of observed peptides (observed MS/MS spectra) in the experiment. Based on these measures of f_i, for differential protein expression was calculated as a Z-score as

$$Z = \frac{f_{i,1} - f_{i,2}}{\sqrt{f_{i,0}(1-f_{i,0})/N_1 + f_{i,0}(1-f_{i,0})/N_2}},$$

where the numerator represents the difference in sampled proportions of protein i in two shotgun proteomics experiments $f_{i,1} = \frac{n_1}{N_1}$, $f_{i,2} = \frac{n_2}{N_2}$; and the denominator represents the standard error of the difference under the null hypothesis in which the two sampled proportions are drawn from the same underlying distribution with the overall

proportion, $f_{i,0} = \frac{n_{i,1} + n_{i,2}}{N_1 + N_2}$. In this analysis, we pooled results from 8 experiments each for INH, EMB, and 5-Cl-PZA treatments and untreated controls.

Clustering of proteins by their expression profiles

Protein expression profiles were calculated from the relative abundance (f_i values) for proteins identified more than 20 times across the 60 shotgun proteomics experiments. Hierarchical clustering was performed on the mean-centered profiles; the Pearson's correlation coefficient between the expression vectors was calculated from the relative abundance (f_i values).

Functional annotation of *M. smegmatis* proteins

The custom *M. smegmatis* predicted protein database was functionally annotated as described in (Wang et al. 2005). The proteins were assigned functional categories with the COG database annotation (Tatusov *et al.* 2001) associated with their top BLAST hits against a database of 89 fully sequenced genomes. *M. smegmatis* proteins were also assigned with *M. tuberculosis* orthologs' GO functional annotation.

We calculated the enrichment of functional categories among proteins showing significant differential expression in drug treatments. Under the hypergeometric distribution, the probability of a functional category being present only at levels expected by random chance is calculated as

$$P_{\text{integrated}}(k \mid n, M, N) = \sum_{i=k}^{\min(M,n)} P_{\text{instantaneous}}(i \mid n, M, N),$$

where, $P_{\text{instantaneous}}(k \mid n, M, N) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$, N is the total population size, M is the number of proteins in the functional category j, n is the number of identified proteins showing significant differential expression in each drug treatment, and k is the overlap between M and n. COG functional categories with $P_{integrated} < (1/18)$ are selected as significant, as there are 18 COG categories in *M. smegmatis*.

Using a two-tailed binomial distribution, we identify GO functional categories showing significant enrichment among proteins differentially expressed in drug treatments as:

$F(k \mid n, p) = \sum_{i=k+1}^{n} (_i^N) p^i (1-p)^{n-i}$, where n represents the total number of the identified proteins, k is the number of identified proteins with the functional category j, p is the fraction of proteins in the functional category j for the whole genome. GO functional categories with $F < (1/350)$ are selected as significant, as there are 350 GO categories. The enrichment is calculated as $f = \frac{n}{N * p}$. For rare GO functional categories with $p < 0.05$, we calculated the significance based on Poisson statistics as:

$F(k \mid n, p) = e^{-p*n}(\sum_{i=0}^{k} \frac{(p*n)^k}{i!})$, where n, k and p are defined as above. Again, GO functional categoies with $F < (1/350)$ are selected as significant.

The mass spectrometry raw data from this study have been deposited in the Open Proteomics Database http://bioinformatics.icmb.utexas.edu/OPD, under accession nos. opd00007_MYCSM–opd00031_MYCSM and opd00047_MYCSM–opd00081_MYCSM.

Supplemental material is available online at http://polaris.icmb.utexas.edu/people/rong/dissertation.

REFERENCES

Aebersold, R. and M. Mann. 2003. Mass spectrometry-based proteomics. *Nature* **422:** 198-207.

Allen, T., P. Shen, L. Samsel, R. Liu, L. Lindahl, and J.M. Zengel. 1999. Phylogenetic analysis of L4-mediated autogenous control of the S10 ribosomal protein operon. *J Bacteriol* **181:** 6124-6132.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Anachkova, B., V. Djeliova, and G. Russev. 2005. Nuclear matrix support of DNA replication. *J Cell Biochem* **96:** 951-961.

Arthur, J.W. and M.R. Wilkins. 2003. Using proteomics to mine genome sequences. *Journal of proteome research* **3:** 393-402.

Banerjee, A., E. Dubnau, A. Quemard, V. Balasubramanian, K.S. Um, T. Wilson, D. Collins, G. de Lisle, and W.R. Jacobs, Jr. 1994. inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis. *Science* **263:** 227-230.

Bennetzen, J.L. and B.D. Hall. 1982. Codon selection in yeast. *J Biol Chem* **257:** 3026-3031.

Betts, J.C., P.T. Lukey, L.C. Robb, R.A. McAdam, and K. Duncan. 2002. Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling. *Mol Microbiol* **43:** 717-731.

Bhat, V.B., M.H. Choi, J.S. Wishnok, and S.R. Tannenbaum. 2005. Comparative plasma proteome analysis of lymphoma-bearing SJL mice. *J Proteome Res* **4:** 1814-1825.

Boshoff, H.I., V. Mizrahi, and C.E. Barry, 3rd. 2002. Effects of pyrazinamide on fatty acid synthesis by whole mycobacterial cells and purified fatty acid synthase I. *J Bacteriol* **184:** 2167-2172.

Boshoff, H.I., T.G. Myers, B.R. Copp, M.R. McNeil, M.A. Wilson, and C.E. Barry, 3rd. 2004. The transcriptional responses of Mycobacterium tuberculosis to inhibitors of metabolism: novel insights into drug mechanisms of action. *J Biol Chem* **279:** 40174-40184.

Brosch, R., A.S. Pym, S.V. Gordon, and S.T. Cole. 2001. The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol* **9:** 452-458.

Bugrim, A., T. Nikolskaya, and Y. Nikolsky. 2004. Early prediction of drug metabolism and toxicity: systems biology approach and modeling. *Drug Discov Today* **9:** 127-135.

Cabusora, L., E. Sutton, A. Fulmer, and C.V. Forst. 2005. Differential network expression during drug and stress response. *Bioinformatics* **21:** 2898-2905.

Carpene, C., S. Bour, V. Visentin, F. Pellati, S. Benvenuti, M.C. Iglesias-Osma, M.J. Garcia-Barrado, and P. Valet. 2005. Amine oxidase substrates for impaired glucose tolerance correction. *J Physiol Biochem* **61:** 405-419.

Chacon, O., Z. Feng, N.B. Harris, N.E. Caceres, L.G. Adams, and R.G. Barletta. 2002. Mycobacterium smegmatis D-Alanine Racemase Mutants Are Not Dependent on D-Alanine for Growth. *Antimicrob Agents Chemother* **46:** 47-54.

Colangeli, R., D. Helb, S. Sridharan, J. Sun, M. Varma-Basil, M.H. Hazbon, R. Harbacheuski, N.J. Megjugorac, W.R. Jacobs, Jr., A. Holzenburg, J.C. Sacchettini, and D. Alland. 2005. The Mycobacterium tuberculosis iniA gene is essential for activity of an efflux pump that confers drug tolerance to both isoniazid and ethambutol. *Mol Microbiol* **55:** 1829-1840.

Cole, S.T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry, 3rd, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, B.G. Barrell, and et al. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393:** 537-544.

Corbett, E.L., C.J. Watt, N. Walker, D. Maher, B.G. Williams, M.C. Raviglione, and C. Dye. 2003. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med* **163:** 1009-1021.

Covert, M.W., E.M. Knight, J.L. Reed, M.J. Herrgard, and B.O. Palsson. 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429:** 92-96.

Dandekar, T., B. Snel, M. Huynen, and P. Bork. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23:** 324-328.

Date, S.V. and E.M. Marcotte. 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol* **21:** 1055-1062.

David, H.L., A. Laszlo, and N. Rastogi. 1989. Mode of action of antimycobacterial drugs. *Acta Leprol* **7 Suppl 1:** 189-194.

de Hoog, C.L. and M. Mann. 2004. Proteomics. *Annu Rev Genomics Hum Genet* **5:** 267-293.

Delcher, A.L., D. Harmon, S. Kasif, O. White, and S.L. Salzberg. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27:** 4636-4641.

Deng, L., K. Mikusova, K.G. Robuck, M. Scherman, P.J. Brennan, and M.R. McNeil. 1995. Recognition of multiple effects of ethambutol on metabolism of mycobacterial cell envelope. *Antimicrob Agents Chemother* **39:** 694-701.

Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95:** 14863-14868.

Eisenberg, D., E.M. Marcotte, I. Xenarios, and T.O. Yeates. 2000. Protein function in the post-genomic era. *Nature* **405:** 823-826.

Eng, J.K., A.L. McCormack, and J.R.3rd. Yates. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5:** 976-989.

Enright, A.J., I. Iliopoulos, N.C. Kyrpides, and C.A. Ouzounis. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402:** 86-90.

Flory, M.R., T.J. Griffin, D. Martin, and R. Aebersold. 2002. Advances in quantitative proteomics using stable isotope tags. *Trends Biotechnol* **20:** S23-29.

Futcher, B., G.I. Latter, P. Monardo, C.S. McLaughlin, and J.I. Garrels. 1999. A sampling of the yeast proteome. *Mol Cell Biol* **19:** 7357-7368.

Gerber, S.A., J. Rush, O. Stemman, M.W. Kirschner, and S.P. Gygi. 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **100:** 6940-6945.

Gorski, S. and T. Misteli. 2005. Systems biology in the cell nucleus. *J Cell Sci* **118:** 4083-4092.

Gygi, S.P., G.L. Corthals, Y. Zhang, Y. Rochon, and R. Aebersold. 2000. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A* **97:** 9390-9395.

Gygi, S.P., B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, and R. Aebersold. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17:** 994-999.

Harth, G., P.C. Zamecnik, J.Y. Tang, D. Tabatadze, and M.A. Horwitz. 2000. Treatment of Mycobacterium tuberculosis with antisense oligonucleotides to glutamine synthetase mRNA inhibits glutamine synthetase activity, formation of the poly-L-glutamate/glutamine cell wall structure, and bacterial replication. *Proc Natl Acad Sci U S A* **97:** 418-423.

Hughes, M.A., J.C. Silva, S.J. Geromanos, and C.A. Townsend. 2006. Quantitative proteomic analysis of drug-induced changes in mycobacteria. *J Proteome Res* **5:** 54-63.

Ishihama, A. 2000. Functional modulation of Escherichia coli RNA polymerase. *Annu Rev Microbiol* **54:** 499-518.

Ishihama, Y., Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, and M. Mann. 2005. Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. *Mol Cell Proteomics* **4:** 1265-1272.

Jacobs, W.R., Jr., G.V. Kalpana, J.D. Cirillo, L. Pascopella, S.B. Snapper, R.A. Udani, W. Jones, R.G. Barletta, and B.R. Bloom. 1991. Genetic systems for mycobacteria. *Methods Enzymol* **204:** 537-555.

Jaffe, J.D., H.C. Berg, and G.M. Church. 2004. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4:** 59-77.

Janeway, C.A.J., P. Travers, M. Walport, and M.J. Shlomchik. 2005. *Immunobiology*. Garlan Science.

Jansen, R., H.J. Bussemaker, and M. Gerstein. 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res* **31:** 2242-2251.

Jolliffe, I.T. 2002. *Principal component analysis*. Springer.

Jungblut, P.R., E.C. Muller, J. Mattow, and S.H. Kaufmann. 2001. Proteomics reveals open reading frames in Mycobacterium tuberculosis H37Rv not predicted by genomics. *Infect Immun* **69:** 5905-5907.

Jungblut, P.R., U.E. Schaible, H.J. Mollenkopf, U. Zimny-Arndt, B. Raupach, J. Mattow, P. Halada, S. Lamer, K. Hagens, and S.H. Kaufmann. 1999. Comparative proteome analysis of Mycobacterium tuberculosis and Mycobacterium bovis BCG strains: towards functional genomics of microbial pathogens. *Mol Microbiol* **33:** 1103-1117.

Kanehisa, M. and S. Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27-30.

Keller, A., A.I. Nesvizhskii, E. Kolker, and R. Aebersold. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74:** 5383-5392.

Kormanec, J., B. Sevcikova, N. Halgasova, R. Knirschova, and B. Rezuchova. 2000. Identification and transcriptional characterization of the gene encoding the stress-response sigma factor sigma(H) in streptomyces coelicolor A3(2). *FEMS Microbiol Lett* **189:** 31-38.

Krude, T., C. Musahl, R.A. Laskey, and R. Knippers. 1996. Human replication proteins hCdc21, hCdc46 and P1Mcm3 bind chromatin uniformly before S-phase and are displaced locally during DNA replication. *J Cell Sci* **109 (Pt 2):** 309-318.

Lea, N.C., S.J. Orr, K. Stoeber, G.H. Williams, E.W. Lam, M.A. Ibrahim, G.J. Mufti, and N.S. Thomas. 2003. Commitment point during G0-->G1 that controls entry into the cell cycle. *Mol Cell Biol* **23:** 2351-2361.

Lee, I., S.V. Date, A.T. Adai, and E.M. Marcotte. 2004. A probabilistic functional network of yeast genes. *Science* **306:** 1555-1558.

Link, A.J., J. Eng, D.M. Schieltz, E. Carmack, G.J. Mize, D.R. Morris, B.M. Garvik, and J.R. Yates, 3rd. 1999. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **17:** 676-682.

Liu, H., R.G. Sadygov, and J.R. Yates, 3rd. 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76:** 4193-4201.

Lu, P., C. Vogel, and E.M. Marcotte. 2005. An estimate of relative contributions of transcriptional and translational regulation by absolute protein expression profiling. *submitted*.

Marcotte, E.M. 2000. Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol* **10:** 359-365.

Marcotte, E.M., M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285:** 751-753.

Marcotte, E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* **402:** 83-86.

Mattow, J., U.E. Schaible, F. Schmidt, K. Hagens, F. Siejak, G. Brestrich, G. Haeselbarth, E.C. Muller, P.R. Jungblut, and S.H. Kaufmann. 2003. Comparative proteome analysis of culture supernatant proteins from virulent Mycobacterium tuberculosis H37Rv and attenuated M. bovis BCG Copenhagen. *Electrophoresis* **24:** 3405-3420.

Mawuenyega, K.G., C.V. Forst, K.M. Dobos, J.T. Belisle, J. Chen, E.M. Bradbury, A.R. Bradbury, and X. Chen. 2005. Mycobacterium tuberculosis functional network analysis by global subcellular protein profiling. *Mol Biol Cell* **16:** 396-404.

Mdluli, K., R.A. Slayden, Y. Zhu, S. Ramaswamy, X. Pan, D. Mead, D.D. Crane, J.M. Musser, and C.E. Barry, 3rd. 1998. Inhibition of a Mycobacterium tuberculosis beta-ketoacyl ACP synthase by isoniazid. *Science* **280:** 1607-1610.

Mikusova, K., R.A. Slayden, G.S. Besra, and P.J. Brennan. 1995. Biogenesis of the mycobacterial cell wall and the site of action of ethambutol. *Antimicrob Agents Chemother* **39:** 2484-2489.

Mitulovic, G., C. Stingl, M. Smoluch, R. Swart, J.P. Chervet, I. Steinmacher, C. Gerner, and K. Mechtler. 2004. Automated, on-line two-dimensional nano liquid chromatography tandem mass spectrometry for rapid analysis of complex protein digests. *Proteomics* **4:** 2545-2557.

Mukhopadhyay, B. and E. Purwantini. 2000. Pyruvate carboxylase from Mycobacterium smegmatis: stabilization, rapid purification, molecular and biochemical characterization and regulation of the cellular level. *Biochim Biophys Acta* **1475:** 191-206.

Nesvizhskii, A.I., A. Keller, E. Kolker, and R. Aebersold. 2003. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75:** 4646-4658.

Nickerson, J. 2001. Experimental observations of a nuclear matrix. *J Cell Sci* **114:** 463-474.

Oda, Y., K. Huang, F.R. Cross, D. Cowburn, and B.T. Chait. 1999. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci U S A* **96:** 6591-6596.

Ong, S.E., B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, and M. Mann. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1:** 376-386.

Ong, S.E. and M. Mann. 2005. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* **1:** 252-262.

Overbeek, R., M. Fonstein, M. D'Souza, G.D. Pusch, and N. Maltsev. 1999. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96:** 2896-2901.

Pandey, A. and M. Mann. 2000. Proteomics to study genes and genomes. *Nature* **405:** 837-846.

Parish, T. and N.G. Stoker. 1998. *Mycobacterial Protocols*. Humana Press.Humana press.

Parish, T. and N.G. Stoker. 2001. *Mycobacterium tuberculosis Protocols.* Humana press.

Pellegrini, M., E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96:** 4285-4288.

Peng, J., J.E. Elias, C.C. Thoreen, L.J. Licklider, and S.P. Gygi. 2003. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* **2:** 43-50.

Perkins, D.N., D.J. Pappin, D.M. Creasy, and J.S. Cottrell. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20:** 3551-3567.

Phetsuksiri B, B.A., Cooper AM, Minnikin DE, Douglas JD, Besra GS, Brennan PJ. 1999. Antimycobacterial activities of isoxyl and new derivatives through the inhibition of mycolic acid synthesis. *Antimicrob Agents Chemother.* **43:** 1042-1051.

Primm, T.P., S.J. Andersen, V. Mizrahi, D. Avarbock, H. Rubin, and C.E. Barry, 3rd. 2000. The stringent response of Mycobacterium tuberculosis is required for long-term survival. *J Bacteriol* **182:** 4889-4898.

Prince, J.T., M.W. Carlson, R. Wang, P. Lu, and E.M. Marcotte. 2004. The need for a public proteomics repository. *Nat Biotechnol* **22:** 471-472.

Radmacher, E., K.C. Stansen, G.S. Besra, L.J. Alderwick, W.N. Maughan, G. Hollweg, H. Sahm, V.F. Wendisch, and L. Eggeling. 2005. Ethambutol, a cell wall inhibitor of Mycobacterium tuberculosis, elicits L-glutamate efflux of Corynebacterium glutamicum. *Microbiology* **151:** 1359-1368.

Raman, K., P. Rajagopalan, and N. Chandra. 2005. Flux balance analysis of mycolic Acid pathway: targets for anti-tubercular drugs. *PLoS Comput Biol* **1:** e46.

Rapaport, E., A. Levina, V. Metelev, and P.C. Zamecnik. 1996. Antimycobacterial activities of antisense oligodeoxynucleotide phosphorothioates in drug-resistant strains. *Proc Natl Acad Sci U S A* **93:** 709-713.

Ratledge, C.a.D., J. 1999. *Mycobacteria molecular biology and virulence*. Blackwell science.

Raychaudhuri, S., M. Basu, and N.C. Mandal. 1998. Glutamate and cyclic AMP regulate the expression of galactokinase in Mycobacterium smegmatis. *Microbiology* **144 (Pt 8):** 2131-2140.

Rison, S.C., S.A. Teichmann, and J.M. Thornton. 2002. Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in Escherichia coli. *J Mol Biol* **318:** 911-932.

Saito, K., S. Fujigaki, M.P. Heyes, K. Shibata, M. Takemura, H. Fujii, H. Wada, A. Noma, and M. Seishima. 2000. Mechanism of increases in L-kynurenine and quinolinic acid in renal insufficiency. *Am J Physiol Renal Physiol* **279:** F565-572.

Salgado, H., G. Moreno-Hagelsieb, T.F. Smith, and J. Collado-Vides. 2000. Operons in Escherichia coli: genomic analyses and predictions. *Proc Natl Acad Sci U S A* **97:** 6652-6657.

Schmidt, F., S. Donahoe, K. Hagens, J. Mattow, U.E. Schaible, S.H. Kaufmann, R. Aebersold, and P.R. Jungblut. 2004. Complementary analysis of the Mycobacterium tuberculosis proteome by two-dimensional electrophoresis and isotope-coded affinity tag technology. *Mol Cell Proteomics* **3:** 24-42.

Scorpio, A. and Y. Zhang. 1996. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nat Med* **2:** 662-667.

Sharp, P.M. and W.H. Li. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15:** 1281-1295.

Shibata, K., T. Fukuwatari, and E. Sugimoto. 2001. Effects of dietary pyrazinamide, an antituberculosis agent, on the metabolism of tryptophan to niacin and of tryptophan to serotonin in rats. *Biosci Biotechnol Biochem* **65:** 1339-1346.

Simpson, J.C., R. Wellenreuther, A. Poustka, R. Pepperkok, and S. Wiemann. 2000. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep* **1:** 287-292.

Sinha, S., K. Kosalai, S. Arora, A. Namane, P. Sharma, A.N. Gaikwad, P. Brodin, and S.T. Cole. 2005. Immunogenic membrane-associated proteins of Mycobacterium tuberculosis revealed by proteomics. *Microbiology* **151:** 2411-2419.

Stein, G.S., J.B. Lian, A.J. van Wijnen, J.L. Stein, A. Javed, M. Montecino, S.K. Zaidi, D. Young, J.Y. Choi, S. Gutierrez, and S. Pockwinse. 2004. Nuclear microenvironments support assembly and organization of the transcriptional regulatory machinery for cell proliferation and differentiation. *J Cell Biochem* **91:** 287-302.

Tabb, D.L., McDonald, W.H., and Yates III, J.R. 2002. DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. 1: 21–26.

Takayama, K. and J.O. Kilburn. 1989. Inhibition of synthesis of arabinogalactan by ethambutol in Mycobacterium smegmatis. *Antimicrob Agents Chemother* **33:** 1493-1499.

Takayama, K., C. Wang, and G.S. Besra. 2005. Pathway to synthesis and processing of mycolic acids in Mycobacterium tuberculosis. *Clin Microbiol Rev* **18:** 81-101.

Talaat, A.M., S.T. Howard, W.t. Hale, R. Lyons, H. Garner, and S.A. Johnston. 2002. Genomic DNA standards for gene expression profiling in Mycobacterium tuberculosis. *Nucleic Acids Res* **30:** e104.

Tatusov, R.L., D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, and E.V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29:** 22-28.

Thackray, P.D. and A. Moir. 2003. SigM, an extracytoplasmic function sigma factor of Bacillus subtilis, is activated in response to cell wall antibiotics, ethanol, heat, acid, and superoxide stress. *J Bacteriol* **185:** 3491-3498.

Tringe, S.G., C. von Mering, A. Kobayashi, A.A. Salamov, K. Chen, H.W. Chang, M. Podar, J.M. Short, E.J. Mathur, J.C. Detter, P. Bork, P. Hugenholtz, and E.M. Rubin. 2005. Comparative metagenomics of microbial communities. *Science* **308:** 554-557.

Vitale, G., R. Pellizzari, C. Recchi, G. Napolitani, M. Mock, and C. Montecucco. 1998. Anthrax lethal factor cleaves the N-terminus of MAPKKs and induces tyrosine/threonine phosphorylation of MAPKs in cultured macrophages. *Biochem Biophys Res Commun* **248:** 706-711.

Wade, M.M. and Y. Zhang. 2004. Anaerobic incubation conditions enhance pyrazinamide activity against Mycobacterium tuberculosis. *J Med Microbiol* **53:** 769-773.

Wang, R., J.T. Prince, and E.M. Marcotte. 2005. Mass spectrometry of the M. smegmatis proteome: protein expression levels correlate with function, operons, and codon bias. *Genome Res* **15:** 1118-1126.

Washburn, M.P., D. Wolters, and J.R. Yates, 3rd. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19:** 242-247.

Washburn, M.P. and J.R. Yates, 3rd. 2000. Analysis of the microbial proteome. *Curr Opin Microbiol* **3:** 292-297.

Wilson, M., J. DeRisi, H.H. Kristensen, P. Imboden, S. Rane, P.O. Brown, and G.K. Schoolnik. 1999. Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization. *Proc Natl Acad Sci U S A* **96:** 12833-12838.

Wu, S., S.T. Howard, D.L. Lakey, A. Kipnis, B. Samten, H. Safi, V. Gruppo, B. Wizel, H. Shams, R.J. Basaraba, I.M. Orme, and P.F. Barnes. 2004. The principal sigma factor sigA mediates enhanced growth of Mycobacterium tuberculosis in vivo. *Mol Microbiol* **51:** 1551-1562.

Yeung, K.Y. and W.L. Ruzzo. 2001. Principal component analysis for clustering gene expression data. *Bioinformatics* **17:** 763-774.

Yi, E.C., M. Marelli, H. Lee, S.O. Purvine, R. Aebersold, J.D. Aitchison, and D.R. Goodlett. 2002. Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis* **23:** 3205-3216.

Yu, J., L. Hederstedt, and P.J. Piggot. 1995. The cytochrome bc complex (menaquinone:cytochrome c reductase) in Bacillus subtilis has a nontraditional subunit organization. *J Bacteriol* **177:** 6751-6760.

Zahrt, T.C. and V. Deretic. 2000. An essential two-component signal transduction system in Mycobacterium tuberculosis. *J Bacteriol* **182:** 3832-3838.

Zhang, Y. 2005. The magic bullets and tuberculosis drug targets. *Annu Rev Pharmacol Toxicol* **45:** 529-564.

Zhang, Y. and D. Mitchison. 2003. The curious characteristics of pyrazinamide: a review. *Int J Tuberc Lung Dis* **7:** 6-21.

Zimhony, O., J.S. Cox, J.T. Welch, C. Vilcheze, and W.R. Jacobs, Jr. 2000. Pyrazinamide inhibits the eukaryotic-like fatty acid synthetase I (FASI) of Mycobacterium tuberculosis. *Nat Med* **6:** 1043-1047.

Chapter 4. A novel metabolic pathway in mycobacteria

As numerous genomes have been fully sequenced in recent years, determining protein functions using genome sequences is the next target. The protein function, instead of the action of a single molecule, is the context of its interaction with other proteins in the cell (Eisenberg *et al.* 2000). We set up a system to predict protein interactions, construct new functional networks, and identify protein pathways (Marcotte et al. 1999b). The method involves three steps. First, application of several computational methods for detecting functional linkages to all pairs of genes. Second, combining all derived functional links and build a functional network from acting pairs of proteins. Third, map known gene functions onto the network and predict their metabolic pathways (Dandekar et al. 1998).

This study was an exercise in assigning function to an uncharacterized pathway by purely computational approaches. An uncharacterized metabolic pathway in *Mycobacterium tuberculosis* that involves seven proteins (Figure 4.1) had previously been identified and was chosen as the subject of the study. It is known that Rv3741c and Rv3742c are probable oxygenase A and B subunits, respectively, Rv3848 is a probable membrane protein, what functions other proteins have and what metabolic pathway these proteins belong to are unknown. In order to detect functional linkages between these proteins, four computational methods were used in our research: phylogenetic profile method (Date and Marcotte 2003; Pellegrini et al. 1999), Rosetta Stone method (Enright et al. 1999; Marcotte et al. 1999a), gene neighbor method (Dandekar et al. 1998; Overbeek et al. 1999), and operon prediction method (Rison et al. 2002; Salgado et al. 2000). These

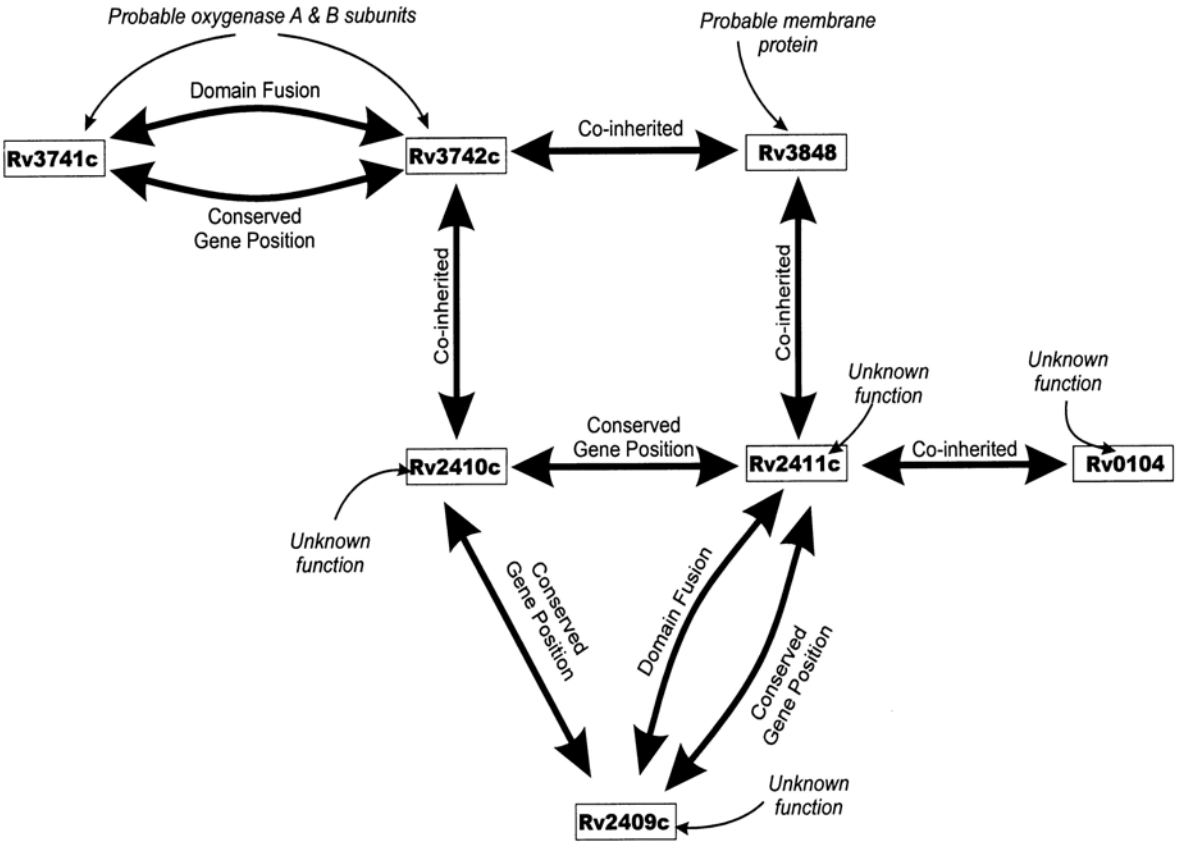


Figure 4.1 The predicted metabolic pathway with seven *M. tuberculosis* proteins.

methods discover functional links between non-homologous proteins, reconstructing pathways through the use of genome sequence data. These methods were integrated with BLASTP (Altschul et al. 1990) based on the protein homology and structure-derived protein functions to predict protein-protein interactions, functional relationships, and metabolic pathways. We plotted a metabolic pathway in *Mycobacterium tuberculosis* and assigned protein functions to the pathway components (Figure 4.2). This pathway might perform unusual amino acids synthesis, carbohydrate and lipid or cell wall metabolism; homologs of several components of this pathway were found in *M. tuberculosis* and other organisms, therefore we predicted parallel pathways exist and have similar functions.

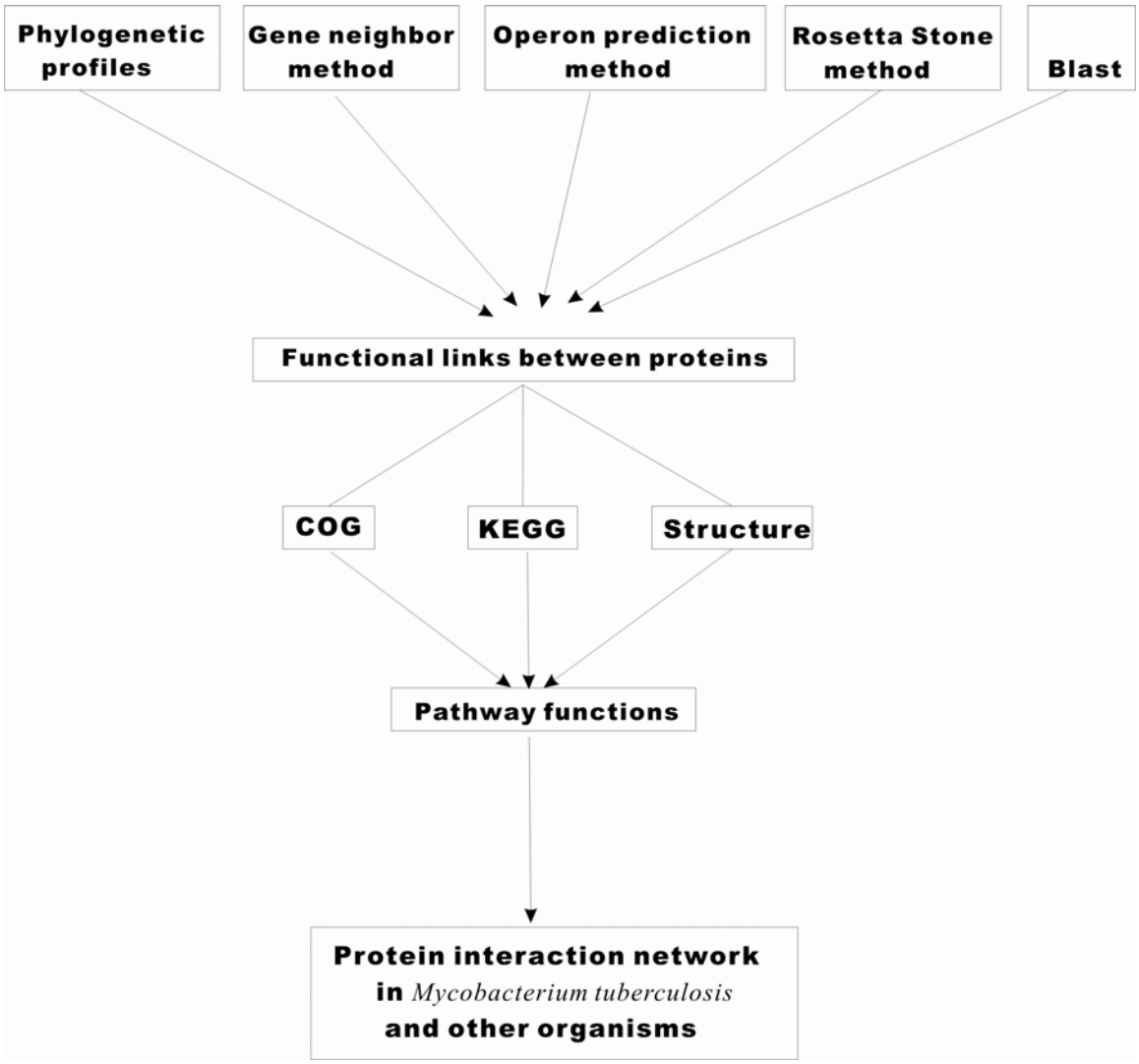


Figure 4.2 The strategy used in predicting protein interactions and pathway functions.

RESULTS

Protein functional links

M. tuberculosis homologs for each protein were identified (Figure 4.2). For example, Rv2565 is homolog of Rv0104; Rv2566 and Rv2569c are homologs of Rv2409c; Rv2567 is homolog of Rv2410c and Rv2411c; Rv0565, Rv3083, and Rv3854c are homologs of Rv3741c and Rv3742c. There are functional links between these homologous proteins. The phylogenetic profile method can also be used in predicting protein-protein interactions, but it rests on entirely different assumptions from BLASTP (Pellegrini et al. 1999). A phylogenetic profile describes the pattern of presence or absence of a particular protein across a set of genomes. Proteins that operate together in the cell are inherited in a correlated fashion. If two proteins have the same phylogenetic profiles, it is inferred that two proteins are co-inherited and might have a functional link (Date and Marcotte 2003; Marcotte 2000; Marcotte et al. 1999b). The similarity between phylogenetic profiles for proteins can be represented by the mutual information scores. In Table 4.1, proteins with mutual information scores above 0.4 show they have similar phylogenetic profiles and indicates that they are functionally linked.

The operon prediction method predicts whether or not a pair of neighboring genes are in the same operon by measuring the physical distance between the gene pair. Neighboring genes were found to be in the same operon with high probability if the distance between them is less than 40 nucleotides (Salgado *et al.* 2000). The result shows that the genes encoding proteins Rv2409c, Rv2410c, Rv2411c and genes Rv3740c, Rv3741c, Rv3742c form two operons, as all consecutive genes are separated by less than 40 nucleotides. The genes encoding Rv3848 and Rv0104 are in different operons, both of

	Rv2409c	Rv2410c	Rv2411c	Rv2566	Rv2567	Rv2569c	Rv3740c	Rv3741c	Rv3742c	Rv0565	Rv3083	Rv3854c	Rv3848	Rv0104	Rv2565
Rv2409c		0.44	0.59	0.41	0.59		0.07	0.17	0.21	0.22	0.23	0.24	0.11	0.07	0.09
Rv2410c	0.44		0.48	0.53		0.48	0.07	0.18	0.22	0.24	0.25	0.26	0.13	0.09	0.09
Rv2411c	0.59	0.48		0.44		0.63	0.06	0.16	0.2	0.19	0.2	0.21	0.11	0.08	0.05
Rv2566	0.41	0.53	0.44		0.53		0.08	0.19	0.24	0.28	0.29	0.3	0.11	0.1	0.09
Rv2567	0.59			0.53		0.63	0.07	0.18	0.22	0.24	0.25	0.26	0.13	0.09	0.1
Rv2569c		0.48	0.63		0.63		0.06	0.16	0.2	0.19	0.2	0.21	0.11	0.08	0.05
Rv3740c	0.07	0.07	0.06	0.08	0.07	0.06		0.1	0.11	0.2	0.21	0.21	0.08	0.12	0.03
Rv3741c	0.17	0.18	0.16	0.19	0.18	0.16	0.1		0.4				0.09	0.11	0.04
Rv3742c	0.21	0.22	0.2	0.24	0.22	0.2	0.11	0.4					0.11	0.13	0.04
Rv0565	0.22	0.24	0.19	0.28	0.24	0.19	0.2						0.11	0.14	0.11
Rv3083	0.23	0.25	0.2	0.29	0.25	0.2	0.21						0.13	0.14	0.11
Rv3854c	0.24	0.26	0.21	0.3	0.26	0.21	0.21						0.12	0.43	0.09
Rv3848	0.11	0.13	0.11	0.11	0.13	0.11	0.08	0.09	0.11	0.11	0.13	0.12		0.16	0.19
Rv0104	0.07	0.09	0.08	0.1	0.09	0.08	0.12	0.11	0.13	0.14	0.14	0.43	0.16		
Rv2565	0.09	0.09	0.05	0.09	0.1	0.05	0.03	0.04	0.04	0.11	0.11	0.09	0.19		
1pqC								0.21	.26**						
1pqP								.22*	0.22						

Table 4.1 Phylogenetic profiles indicate functional links between pathway members.

Each value represents the mutual information measurement of the similarity between two phylogenetic profiles. Values above ~0.4 indicate two genes are functionally linked (cyan boxes). Stars represent highest scores for Rv3741c and Rv3742c to non-monooxygenases in *M. tuberculosis* geneome; yellow boxes show the second highest score for Rv3742c. Both stars and yellow boxes indicate functional linkage.

which are composed of single gene. We predict that these proteins in the same operon have similar functions given that genes in a same operon tend to have similar functions or in the similar cellular pathway (Marcotte 2000; Salgado *et al.* 2000). In addition, the shorter the pathway distance (the number of distinct metabolic steps separating two enzymes), the smaller the gene intervals. For example, enzymes coded by nearby genes in *E. coli* genome are more likely than distant ones to be close in a biochemical pathway (Rison et al. 2002). Therefore, we predict that proteins encoded by the Rv2409c operon (including Rv2409c, Rv2410c, and Rv2411c) are close in a cellular pathway, as are the proteins encoded by Rv3740c operon (Rv3740c, Rv3741c, and Rv3742c). Furthermore, proteins in each operon may have similar physiological functions as well.

In the Rosetta Stone method (Enright et al. 1999; Marcotte et al. 1999a), separate proteins A and B in one organism are sometimes expressed as a fusion protein in another species, implying proteins A and B might be linked in physiological function. In this manner, Rv2567 is a homolog of Rv2410c and Rv2411c; Rv0565, Rv3083, and Rv3854c are homologs of Rv3741c and Rv3742c. Therefore, we predict Rv2410c and Rv2411c are functionally linked by domain fusion, same as Rv3741c and Rv3742c. In the gene neighbor method (Dandekar et al. 1998; Overbeek et al. 1999), if the genes that encode two proteins are neighbors in the genome of other organisms, they tend to be functionally linked. Proteins Rv2409c, Rv2410c and Rv2411c and proteins Rv3740c, Rv3741c, and Rv3742c are connected within each group by conserved gene positions. Rv0104 and Rv2409c are functionally linked in this method as their homologs Rv2565 and Rv2566 are encoded in the same operon.

Construction of novel pathways

Taking results of individual functional links from the methods of BLASTP, operon prediction, phylogenetic profiles, Rosetta Stone, and gene neighbors, we combine a set of protein interactions into an extended cellular pathway (Figure 4.3). (1) Proteins Rv2409c, Rv2410c, and Rv2411c are functionally linked by conserved gene position, domain fusion, and operon prediction, as are proteins Rv3740c, Rv3741c, and Rv3742c. (2) Protein Rv2409c is connected to Rv0104 since they have conserved gene neighbors. (3) The group of Rv2409c, Rv2410c, and Rv2411c have similar phylogenetic profiles as group of Rv2566, Rv2567, and Rv2569c, as indicated by the high mutual information scores, which infers that these proteins are parallel pathways in *M. tuberculosis*. We found homologs of several components in *M. tuberculosis* and other organisms, so we infer that this pathway exists at least partially in these organisms, such as *Caulobacter crescentus*, *Deinococcus radiodurans, Mycobacterium leprae,* and *Pseudomonas aerufinosa*.

Identification of the pathway functions

Having demonstrated that these proteins are functionally linked, we want to infer the biological functions of these proteins. Structure motifs and evolutionary tracing provide functional information for proteins and which protein superfamily and structural model each protein belongs to (Table 4.2). Table 4.3 shows enzymatic functions for proteins predicted from the COG database (Tatusov et al. 2001). In Table 4.4, the metabolic pathways (KEGG pathway)(Kanehisa and Goto 2000) in which these proteins might be performing are consistent with predicted protein functions. It is known that

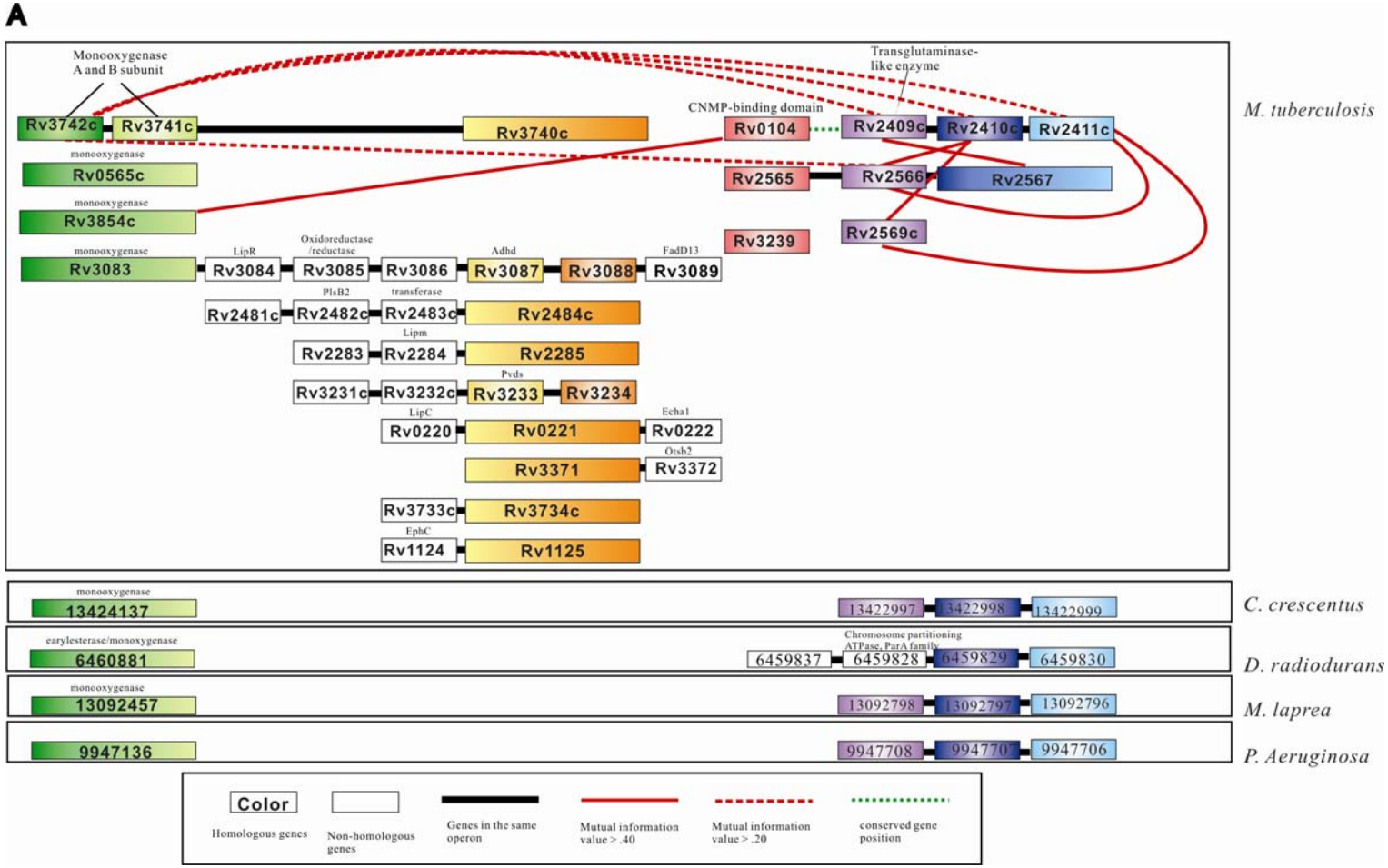

A
Monooxygenase
A and B subunit
Transglutaminase-
like enzyme
CNMP-binding domain
Rv3742c
Rv3741c
Rv3740c
Rv0104
Rv2409c
Rv2410c
Rv2411c
M. tuberculosis
monooxygenase
Rv0565c
Rv2565
Rv2566
Rv2567
monooxygenase
Rv3854c
Rv2569c
Rv3239
monooxygenase
Rv3083
LipR
Rv3084
Oxidoreductase
/reductase
Rv3085
Rv3086
Adhd
Rv3087
Rv3088
FadD13
Rv3089
PlsB2
transferase
Rv2481c
Rv2482c
Rv2483c
Rv2484c
Lipm
Rv2283
Rv2284
Rv2285
Pvds
Rv3231c
Rv3232c
Rv3233
Rv3234
LipC
Echa1
Rv0220
Rv0221
Rv0222
Otsb2
Rv3371
Rv3372
Rv3733c
Rv3734c
EphC
Rv1124
Rv1125
monooxygenase
13424137
13422997
3422998
13422999
C. crescentus
earylesterase/monoxygenase
6460881
Chromosome partitioning
ATPase, ParA family
6459837
6459828
6459829
6459830
D. radiodurans
monooxygenase
13092457
13092798
13092797
13092796
M. laprea
9947136
9947708
9947707
9947706
P. Aeruginosa
Color
Homologous genes
Non-homologous
genes
Genes in the same
operon
Mutual information
value > .40
Mutual information
value > .20
conserved gene
position

Figure 4.3

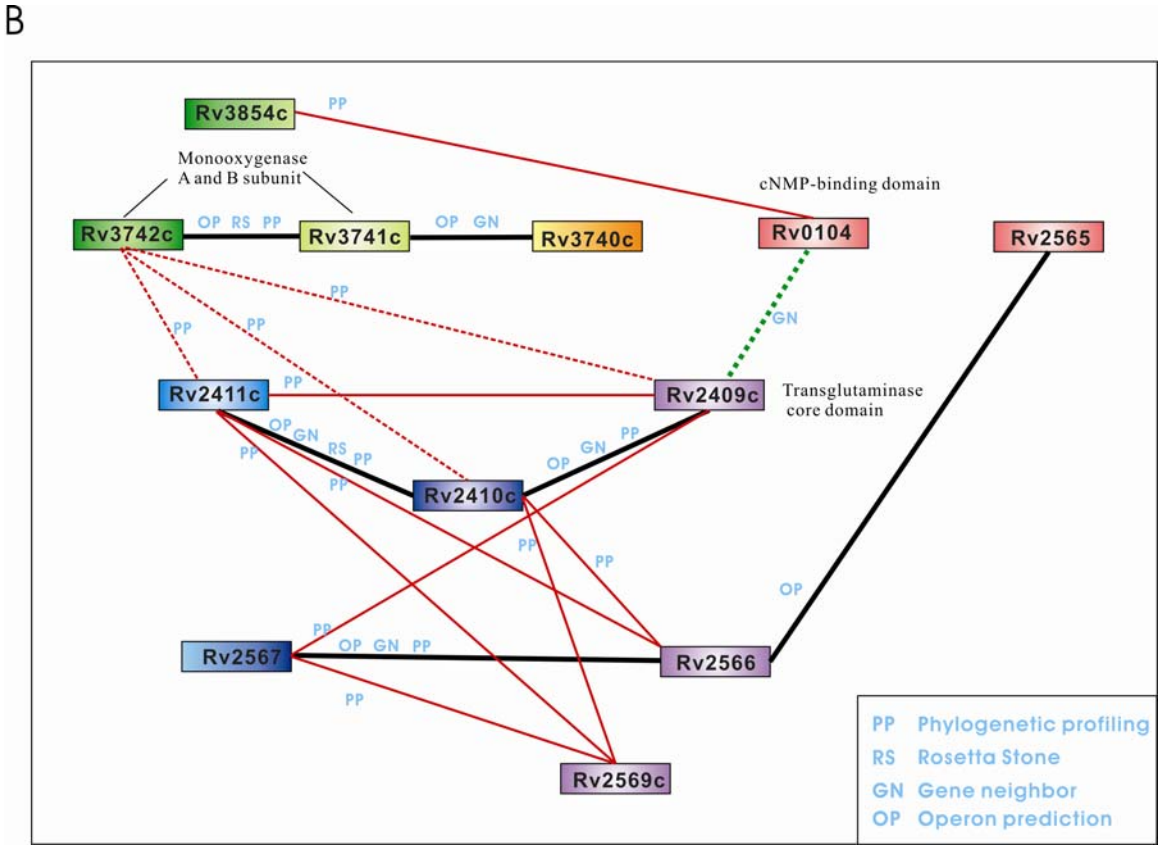


Figure 4.3 Parallel protein interaction pathways predicted in M. tuberculosis and other organisms.

(A) Extended protein pathways in *M. tuberculosis* and other organisms. (B) Core pathways in *M. tuberculosis*. Proteins connected by black bars are encoded by genes in the same operon; proteins in the same-colored boxes are homologous to each other (which are in one column in A), non-colored boxes represent non-homologous proteins. Red lines show the mutual information scores are higher than 0.40; red dotted lines represent top high scores for Rv3742c. Green dotted line indicates that operon Rv2409c is functionally linked to Rv0104 by gene neighbor method. Protein Rv3742c is linked to the Rv2566 since the mutual information score between them is the highest one (0.24) for Rv3742c to non-monooxygenases in *M. tuberculosis* genome. Protein Rv3742c is linked to Rv2409c, Rv2410c, and Rv2411c as the mutual information scores between each pair are high. A closely related pathway in *M. tuberculosis* involves Rv2565, Rv2566, Rv2567, and Rv2569c.

TB protein	Function (from COGs and BLASTP)	Domain
Rv2409c	Predicted intracellular enzyme; amino acid transport and metabolism	transglutaminase
Rv2566	Predicted intracellular enzyme; amino acid transport and metabolism	transglutaminase
Rv2569c	Predicted intracellular enzyme; amino acid transport and metabolism	transglutaminase
Rv3848	Probable membrane protein	
Rv3741c	Similar to probable monooxygenase; inorganic ion transport and metabolism	
Rv3742c	Similar to probable monooxygenase; inorganic ion transport and metabolism	
Rv0565	Probable monooxygenase; inorganic ion transport and metabolism	
Rv3083	Probable monooxygenase; inorganic ion transport and metabolism	
Rv3854c	Probable monooxygenase; inorganic ion transport and metabolism	
Rv0104	Signal transduction	cNMP-binding
Rv2565	Nucleotide binding	cNMP-binding
Rv3239c	Has homology with transmembrane efflux protein; carbohydrate transport and metabolism	cNMP-binding

Table 4.2 Protein functions predicted from the protein superfamilies and structural models.

	Superfamily http://maple.bioc.columbia.edu/pp/submit_meta.html	Structural model http://maple.bioc.columbia.edu/pp/submit_meta.html
Rv2409c	Cysteine proteinasaes superfamily	No
Rv2410c	No	Possible similar to 1sig, crystal structure of A sigma70 subunit fragment from *E. coli* RNA polimerase; 1gal, Glucose oxidase
Rv2411c	Glutathione synthetase ATP-binding domain-like superfamily	No
Rv3740c	No	No
Rv3741c	FAD/NAD(P)-binding domain	No
Rv3742c	FAD/NAD(P)-binding domain	No
Rv0104	cAMP-binding domain-like NAD(P)-binding Rossmann-fold domains, Formate/glycerate dehydrogenase catalytic domain-like	Similar to 1rgs, regulatory subunit of cAMP dependent protein kinase
Rv3848	No	Possible similar to 1tsl, tyrosyl-transfer RNA synthetase

Table 4.3 Protein functions predicted from the COG database.

Candidate pathways	Evidence
Glutathione metabolism	Rv2409c transglutaminase or glutamyltransferase can be involved in glutathione metabolism Rv2411c is a member of glutathione synthetase superfamily
Selenoamino acid metabolism	Transglutaminase or glutamyltransferase(like Rv2409c) involved in such metabolisms
Cyanoamino acid metabolism	
Taurine and hypotaurine metabolism	
Prostaglandin and leukotriene metabolism	
Pentose phosphate metabolism	Rv2410c has sequence homologue with 1gal (Glucose oxidase), which is in pentose phosphate metabolism
Phenylalanine, tyrosine and tryptophan biosynthesis	Rv3848 has predicted structural homology with 2tsl (tyrosyl-transfer tRNA synthetase), which is involved in phenylalanine, tyrosine and tryptophan biosynthesis, aminoacyl-tRNA biosynthesis
Aminoacyl-tRNA biosynthesis	

Table 4.4 *M*etabolic pathways consistent with predicted enzymatic functions of proteins in the pathway.

there is no glutathione metabolism in *M. tuberculosis* (Cole et al. 1998b), therefore protein encoded in the Rv2409c operon and the protein Rv0104 may take part in unus amino acid synthesis, carbohydrate or complex lipid metabolism. A closely rela pathway in *M. tuberculosis* may involve Rv2565, Rv2566, and Rv2567 as well Rv2569c. Rv3741c and Rv3742c, both linked to the Rv2409c system, have sim phylogenetic profiles with lipoproteins lpqP and lpqC, respectively, therefore Rv37 and Rv3742c might be involved in lipid or cell wall metabolism. Finally, we concl that this pathway is involved in the biological pathway of unusual amino acid synthe carbohydrate and complex lipid/cell wall metabolism. Some parallel pathways in *tuberculosis* and other species were also identified. Rison *et al*. suggested that parallel pathways might be chemistry conserved, co-factor substrate, or substrate bind conserved (Rison et al. 2002). Therefore, we propose those parallel pathways in e organism might be functionally linked.

DISCUSSION

The functions and relationships between genes and proteins can be obtained from the genome-wide experiments, for example, mRNA expression in microarray experiment and protein expression in shotgun experiment, or small-scale experiments such as pull-down experiment. Patterns of gene fusion, conservation of gene position, gene co-inheritance, and other types of evolutionary information can also be used to discover protein functions (Eisenberg *et al.* 2000). These computational methods are especially useful when there is insufficient or minimal experimental evidence.

The main goal of our research is to find functions for hypothetical proteins and assign the unknown proteins to the metabolic pathway. The important question is how we predict protein interaction from genome sequence alone, which will speed the research and provide direction for specific proteins. Our combination of computational methods such as phylogenetic profiles, domain fusion, and gene neighbor, provides new information about functional relationships and hence goes beyond the capabilities of traditional sequence matching and makes it possible to predict protein-protein interactions from genome sequences. We began with a simple metabolic pathway with uncharacterized proteins in *M. tuberculosis*. Constructing metabolic pathways from combined computational methods allows the novel prediction of proteins functions that can not be inferred reliably from any single method.

We characterized the metabolic functions of proteins and identified particular pathway functions. We predicted that this metabolic pathway is involved in the biological pathway of unusual amino acid synthesis, carbohydrate metabolism and complex

lipid/cell wall metabolism. Parallel pathways in *M. tuberculosis* and other species were also identified and proposed to be functional linked. Overlapping predictions among methods gives confidence in the methods ability to predict true links. This allows us to confidently predict functions for uncharacterized proteins that are linked to characterized proteins. Proteins are characterized according to different metabolic pathways. These characterized proteins allow further study of the parallel metabolic pathways in *M. tuberculosis* and other organisms.

METHODS

The methods we used to predict protein interactions and functions are shown in Figure 4.2. The methods of BLASTP, operon prediction, phylogenetic profiles, Rosetta Stone, and gene neighbors have been used to predict protein functional links. Homologs of proteins were obtained by comparing the amino acid sequences against the protein sequences nr database (http://www.ncbi.nlm.nih.gov) using the program BLASTP and default parameters, and selecting the protein homolog of the BLASTP match when it surpassed a BLASTP expectation value threshold of 10^{-6}. Phylogenetic profiles were taken from (Date and Marcotte 2003), and mutual information scores between phylogenetic profiles were calculated as described. Operon structures were inferred from the distances between genes available through Entrez Genome (http://www.ncbi.nlm.nih.gov). Protein functions and pathway functions were inferred as follows: information about protein families was obtained from (http://maple.bioc.columbia.edu/pp/submit_meta.html). Functions of proteins were obtained from the COGs website ((Tatusov et al. 2001); http://www.ncbi.nlm.nih.gov/COG). Information about metabolic pathways was obtained from the KEGG pathway database (http://www.genome.admjp/kegg/kegg2.html).

REFERENCES

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Cole, S.T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry, 3rd, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M.A. Quail, M.A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J.E. Sulston, K. Taylor, S. Whitehead, and B.G. Barrell. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393:** 537-544.

Dandekar, T., B. Snel, M. Huynen, and P. Bork. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23:** 324-328.

Date, S.V. and E.M. Marcotte. 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol* **21:** 1055-1062.

Eisenberg, D., E.M. Marcotte, I. Xenarios, and T.O. Yeates. 2000. Protein function in the post-genomic era. *Nature* **405:** 823-826.

Enright, A.J., I. Iliopoulos, N.C. Kyrpides, and C.A. Ouzounis. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402:** 86-90.

Kanehisa, M. and S. Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27-30.

Marcotte, E.M. 2000. Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol* **10:** 359-365.

Marcotte, E.M., M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285:** 751-753.

Marcotte, E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* **402:** 83-86.

Overbeek, R., M. Fonstein, M. D'Souza, G.D. Pusch, and N. Maltsev. 1999. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96:** 2896-2901.

Pellegrini, M., E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96:** 4285-4288.

Rison, S.C., S.A. Teichmann, and J.M. Thornton. 2002. Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in Escherichia coli. *J Mol Biol* **318:** 911-932.

Salgado, H., G. Moreno-Hagelsieb, T.F. Smith, and J. Collado-Vides. 2000. Operons in Escherichia coli: genomic analyses and predictions. *Proc Natl Acad Sci U S A* **97:** 6652-6657.

Tatusov, R.L., D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, and E.V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29:** 22-28.

Chapter 5. Application of LC/LC/MS/MS on Proteome Analysis

Hematopoietic cells have the ability to proliferate from quiescent state in response to antigenic responses. Cells change considerably during this process: the chromatin becomes less dense, the nucleoli appear, the volume of both the nucleus and the cytoplasm increases, and new mRNA and proteins are synthesized (Janeway et al. 2005). Much of the cellular regulatory machinery controlling this process is localized to the nuclear matrix; mediating control of both DNA replication and gene expression by providing a structural support and organizing the many protein factors involved (Anachkova et al. 2005; Stein et al. 2004). The nuclear matrix is a structure in the nucleus, generally consisting of a nuclear lamina and pore complexes surrounding an internal fibrogranular network of RNP proteins and residual nucleoli. It spatially organizes DNA, RNA, and proteins in the nucleus. This is important as distinct macromolecular complexes forming on the nuclear matrix are involved in DNA replication, repair, and recombination, as well as RNA polymerase II transcription, RNA processing and metabolism (Nickerson 2001). In human primary T lymphocytes, it has been found that a DNA origin recognition complex forms when the cells are in late G_1 phase and is spatially organized by being bound to the nuclear matrix (Lea et al. 2003).

The rapidly evolving field of mass spectrometry-based proteomic investigation allows for the identification of novel proteins and the characterization of protein functions in different cellular compartments (Washburn et al. 2001). More than 480 mouse plasma proteins in normal and 790 mouse plasma proteins in RcsX-tumor-bearing SJL mouse plasma were identified using the 1D-Gel-LC-MS/MS method (Bhat et al.

2005). 174 matrix-bound proteins were identified using LC/LC/MS/MS approach in the extracts of the Jurkat cell line (Mitulovic et al. 2004). In our study, LC/LC/MS/MS is used to identify proteins from samples and relative protein abundances are approximated from the rates of sampling by repeated injections of the same sample into the LC/LC/MS/MS (Wang et al. 2005), and the number of peptides identified for each protein as in Chapter 3.

To investigate the mechanisms that control cell cycle transitions from $G_0 \rightarrow G_1$-phase, we carried out protein expression profiling to identify "signatures" of cell cycle phase transitions. Nuclear matrix-associated proteins that are regulated during the $G_0 \rightarrow G_1$-phase transition in human peripheral blood T cells were identified using LC/LC/MS/MS technology. The characterization of these proteins provides the foundation for analyzing the matrix-bound proteome of human primary haemopoietic cells and a comprehensive investigation of regulation mechanisms in the transition through the $G_0 \rightarrow G_1$-phase in haemopoietic cells.

In a second study of mammalian nuclear protein, we report the proteomic characterization of the nuclear protein from mouse T lymphoma. The mammalian nucleus is a highly heterogeneous organelle, the most complex subcellular compartment containing the vast majority of genetic material and the site of all major genome regulatory processes (Gorski and Misteli 2005). It is estimated that 20% of cellular proteins are in the nucleus of mammalian cells (Simpson et al. 2000). We identify 116 proteins by shotgun proteomics and demonstrate several classes of functional proteins located in the nucleus. This study will shed light on how the nucleus functions and provide an understanding of the molecular mechanisms involved in nuclear processes.

RESULTS

Part I. Proteomic analysis of nuclear matrix-associated proteins in primary human T-lymphocyte activation

We investigated the dynamics of the proteins that bind the nuclear matrix by analyzing T lymphocyte samples taken in the quiescent state G_0 phase and in the active growth state G_1 phase. Matrix-associated proteins were extracted by our collaborators at King's College, UK (N. Shaun Thomas's lab) and four different protein mixtures: G_0-matrix-bound, G_0-matrix-free, G_1-matrix-bound, and G_1-matrix-free were prepared. The same biological samples were each analyzed 4 times using LC/LC/MS/MS, and matrix-associated proteins present in primary T cells in G_0 and G_1 phase were identified. In total, 613 proteins have been identified across all 16 shotgun experiments, with the false positive protein identification error rate at less than 2% (Supplemental Table 5.1).

Of 404 matrix-bound proteins, 255 proteins become matrix-bound as cells progress from G_0 to late G_1 phase, while 67 proteins were bound in G_0 that are not bound in G_1 phase. Interestingly, the most enriched-proteins among these G_0-specific proteins are methyl CpG binding protein 2 and lymphocyte-specific protein 1 (pp52). Many abundant matrix-bound proteins were identified: histone proteins (H1, H2A, H2B, H3, and H4) and non-histone proteins (DEAD box helicases enriched in G1-phase), hnRNP (C, U), snRNP (D1, D3, U1, U2, U3, U5, and U4/U6.U5 tri snRNP), nucleolar (*e.g.* NOP56, Nucleolin (enriched in G1-phase, apparently an angiogenesis marker)), and matrix proteins (Lamins, nuclear mitotic apparatus proteins (NUMA)). In all, 183 of the 404 matrix-bound proteins were previously known to be nuclear.

Other interesting G1-enriched matrix-bound proteins include: WD repeat domain 36, T-cell activation WD repeat protein, interleukin enhancer binding factor 2, nuclear factor of activated T-cells, 45-kDa, and nuclear factor NF-κB1. Importantly, a number of low abundance transcription factors were also identified from the samples. These proteins includes several SWI/SNF related transcription factors, transcription factor 3 (AML2, Acute myeloidleukemia 2 protein), proliferation-associated protein 2G4 (induced in G1 and may help control replication), apoptosis antagonizing transcription factor (activates DNA synthesis in quiescent NIH-3T3 cells through HDAC1 displacement), NFAT, TRRAP transformation/transcription domain-associated protein (histone acetylase-related), CCAR1 cell-cycle, apoptosis regulatory protein 1, and so on. These examples demonstrate that by purifying the nuclear matrix, low abundance proteins like transcription factors can be successfully observed using shotgun proteomics.

We also identified splicing factors and transcriptional co-activators. These include splicing factors 1, 2, 3a, 3b and Arg/Ser rich 7 & 10, transcriptional co-activator p52/p75, bZip and n-pac. Some proteins present only in late G_1 phase (including bZip, MCM3, 5, 6 and 7, HDAC2, and DNA topoisomerase I) indicate that we are able to detect components of complexes that are induced or form during cell cycle entry. The identified matrix-associated proteins are involved in facilitating DNA replication, RNA processing, and RNA transport, among other functions.

Matched controls of non-matrix proteins run in parallel accounted for 38% of the proteins identified, although it is unclear at present whether this is due to cross-contamination or that certain proteins are present in both matrix and non-matrix protein

pools. Abundant cytosolic proteins (glycolysis proteins, alpha and beta actin, tubulin, profilin, and heat shock proteins) were correctly identified as non-matrix proteins.

Part II. Proteomic profiles of nucleus in mouse T lymphoma cells

Nuclear proteins were analyzed using 4 sequential LC/LC/MS/MS experiments, and 848 proteins were identified in all these experiments with a false identification rate of less than 5%. Cytosolic proteins (also 4 experiments) were analyzed in parallel as the control for nuclear protein characterization. Proteins were assigned as nuclear using a Z-score based on the number of peptides identified per protein in the nuclear and cytosolic experiments. 116 proteins are nuclear proteins with 99% confidence ($Z \geq 2.33$, single-sided) (Supplemental Table 5.2). Figure 5.1 shows that the higher the Z value, the higher the fraction of the identified proteins previously known to be nuclear proteins, validating this approach.

These nuclear proteins include histone deacetylase, DNA helicase, DNA methyl-transferase, HMGB2, radixin, methyl-CpG binding protein, acidic nuclear phosphoprotein 32 family member B, tumor rejection antigen gp96, and TBP-interacting protein. Importantly, we also identified cell division cycle 2 homolog A, transcription factor Swi, and transcription elongation factor. In addition, we also detected abundant HnRNP proteins (A3, L, M, H2, R, I, U, K, H1), splicing factor 3b, snRNP (A, E, U2, B, D1), ubiquitin-conjugating enzyme, and valyl-tRNA synthetase 2. These results indicate that low abundance proteins such as transcription factors can successfully be identified by shotgun proteomics of the nucleus, although we observe better coverage of such proteins by directly purifying the nuclear matrix, as in the previous section.

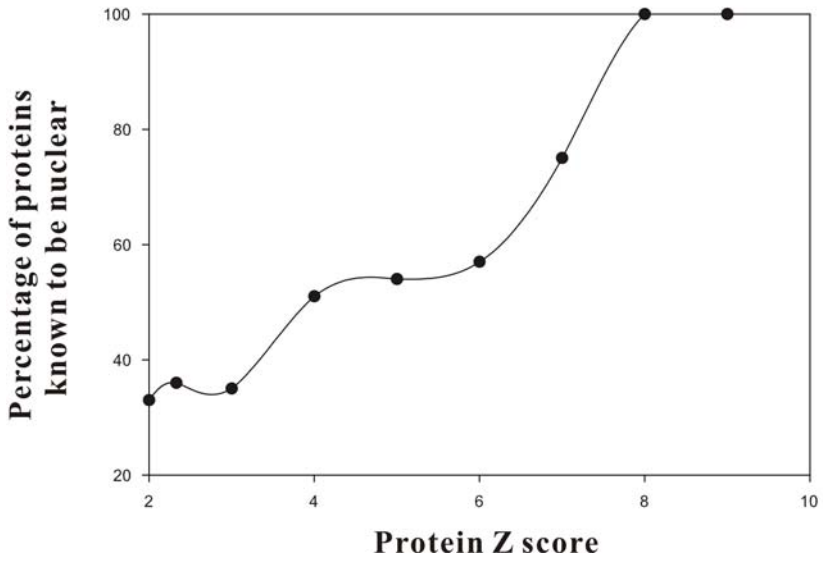


Figure 5.1 Mouse proteins with higher Z scores are enriched for nuclear proteins.

The subcellular localizations for proteins were obtained from the DAVID web server at http://david.niaid.nih.gov/david/version2/index.htm . 36% of the significant proteins ($Z \geq 2.33$) are nuclear proteins; 100% for the proteins ($Z \geq 8$). Proteins with higher Z scores have higher chance to be nuclear proteins, indicating that the Z score effectively identifies nuclear proteins.

Translation initiation factor 5A and translation elongation factor 1 were observed to be abundant in both nucleus and cytosol. Some cytosolic proteins were also identified in the nuclear sample, although it is unclear whether this is due to cross-contamination or that certain proteins are present in both nucleus and cytosolic pools, reflecting their ability to translocate between these compartments. These proteins are ELAV (embryonic lethal abnormal vision), proliferating cell nuclear antigen, endoplasmic reticulum protein Pdia3, zinc-finger proteins, matrin, septin, actin, protein phosphatase (Ppp1ca, Ppp2cb, Anp32b, Pnkp), cytochrome C, protein disulfide isomerase-related protein, vemintin, golgi coil-coiled protein Gcc1, heat shock proteins, dynein, and spectrin. This research justifies a more comprehensive investigation for roles into nuclear protein shotgun proteomics.

DISCUSSION

The mass spectrometer preferentially samples the most abundant proteins; with repeated runs, lower abundant proteins are sampled stochastically, and relative protein abundance can be approximated from the rates of sampling. We can increase the "depth" of the analyses by repeated injections of the same sample into the LC/LC/MS/MS. Repeated injections of the same sample would identify more proteins because the complexity of the sample exceeds the mass spectrometer's capacity for a single analysis. However, the more important point is that many proteins are observed multiple times. The number of observations of a protein is proportional to the protein's absolute abundance (Wang et al. 2005). Therefore, multiple injections not only increase the coverage, they also provide approximate (order-of-magnitude) quantification, and we can detect differential enrichment of proteins from the sampling depth under different conditions.

Shotgun proteomics is a powerful tool for global analysis of nuclear proteins, which provide significant new information on the important regulatory proteins. We apply shotgun proteomics to identify multiple components of nuclear matrix-associated complexes involved in biologically important processes that change through the $G_0 \rightarrow G_1$-phase transition. The proteins identified in our study were predominantly either proteins of the nuclear matrix/cytoskeleton or those known to be associated with the matrix. Three proteins (Nucleolin, HDAC2, and MCM7) have been validated by western blotting (finished by members in Dr. Shaun Thomas' lab). It should be noted that HDAC2 was not known to be induced during transition from G_0 to G_1-phase. These studies set up a platform for analyzing the matrix-bound proteome of human primary haemopoietic cells

and investigating the molecular mechanisms involved in controlling the entry into the cell cycle.

MATERIAL AND METHODS

Database

A database of 27,960 human protein sequences was downloaded from Entrez Genome (http://www.ncbi.nlm.nih.gov). A shuffled version of the database was generated as follows: the amino acid sequences of each predicted protein encoded in the human genome were written in random order, thereby preserving the length and amino acid frequency distribution of each protein but not the amino acid order (as described in Chapter 2).

A database of 25,371 mouse (*Mus musculus*) proteins was downloaded from Entrez Genome (http://www.ncbi.nih.nlm.gov). The subcellular localizations for proteins were obtained from the DAVID web server at http://david.niaid.nih.gov/david/version2/index.htm .

Human primary T cells nuclear matrix-associated protein complexes preparation (performed by members of Dr. Shawn Thomas' lab, King's College, London)

Approximately $3x10^7$ quiescent human primary T cells were isolated from blood. Half of the cells were stimulated with anti-CD3/CD28 for 8 hours and collected 40 hours after the stimulation (Lea et al. 2003), so they enter mid/late G_1 phase and the unstimulated cells remain in G_0 phase. The matrix-associated proteins in the quiescent (G_0) T cells and those in mid/late G_1 T cells were extracted in ice-cold CSK buffer (10 mM PIPES pH 6.8, 100 mM NaCl, 300 mM sucrose, 1 mM $MgCl_2$, 1 mM EDTA, 0.1% (v/v) Triton X-100, 0.1 mM ATP with cocktails of protease and phosphatase inhibitors) (Krude et al. 1996). The soluble (matrix-free) proteins were retained and the insoluble

nuclear matrix/cytoskeleton-associated proteins were washed twice using ice-cold CSK buffer.

Mouse T lymphoma cells nuclear protein sample preparation (performed by Xin Yao, laboratory of Dr. Philip Tucker, UT-Austin)

Approximately $3x10^7$ mouse T lymphoma BW5147 cells were harvested, washed in PBS, and resuspended in 5ml buffer (10mM Hepes pH7.9, 1.5 mM $MgCl_2$, 10mM KCl, 1mM DTT, protease inhibitors) for 10 min. Cells were pelleted (1,000g for 10 min) and resuspended in 2ml of the same buffer. Cells were then lysed using 10 strokes in a homogenizer and nuclei were pelleted (1,000g for 10 min), the supernatant was retained as the cytoplasmic protein sample. The wash and centrifugation steps were repeated once, centrifuging at 30,000g for 20 min. Nuclei were resuspended in 1ml buffer (20 mM Hepes pH7.9, 25% glycerol 0.42 M NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 1 mM DTT, protease inhibitors), and homogenized (~30 strokes), and stirred on a magnetic stirrer for 30-60 min. The lysed nuclei were centrifuged at 30,000g for 20 min) and the supernatant dialyzed against 150 volumes of buffer (20mM Hepes pH7.9, 20% glycerol, 100mMKCl, 0.2mM EDTA, 1mM DTT, protease inhibitors) for 3-4 hours. Nuclear extracts were centrifuged (30,000g for 20 min) and the supernatants collected for analysis.

LC/LC/MS/MS analysis

The protein mixtures were diluted in digestion buffer (50mM Tris HCL pH8.0, 1.0M Urea, 2.0mM $CaCl_2$), trypsinized, and analyzed by LC/LC/MS/MS as described in Chapter 3. Gas phase fractionation (GPF) was used to achieve maximum proteome coverage and increase coverage of low abundant proteins (Yi et al. 2002). Three

sequential LC/LC/MS/MS analyses were performed. Spectra from different mass/charge (*m/z*) ranges (300–680, 680–900, and 900–1500 *m/z*) were collected for data-dependent precursor ion selection; for each MS spectra, the 5 tallest individual peaks, corresponding to peptides, were fragmented by collision-induced dissociation with helium gas to produce MS/MS spectra. Fragmentation data from the three runs were combined for computational analysis.

Protein identification

For the human primary T lymphocyte samples, MS/MS fragmentation spectra were analyzed using the program TurboSequest/ BioWorks 3.1 and DTASelect software (selection criteria are ≥ 2 unique peptides/protein; Xcorr ≥ 1.8, ≥ 2.5 & ≥ 3.5 for +1, +2 & +3 charged peptides). The false positive identification rate was estimated by searching fragmentation spectra against the shuffled human protein database (Wang et al. 2005). By this analysis, DTASelect gives a low false positive identification rate (<2%), presumably with a high false negative identification rate. We estimated protein abundances from the number of observations of each protein. Proteins enriched in the nuclear matrix were identified by testing if the total number of observations in matrix-bound samples minus the total number of observations in matrix-free samples was greater than 0, indicating that the proteins are enriched in the matrix bound fractions relative to the matrix-free ones. Likewise, proteins enriched in G_1-phase relative to G_0 were identified by requiring the total number of observations in G_1-bound samples minus the total number of observations in G_0-bound samples to be greater than 0.

For mouse T lymphoma nuclear protein complex samples, MS/MS fragmentation spectra were analyzed using the program TurboSequest/ BioWorks 3.1 and

ProteinProphet (Nesvizhskii et al. 2003). The numbers of identified peptides per protein were used to estimate the relative enrichment of the proteins in the nucleus, as described in Chaper 3.

Supplemental data is available on line at
http://polaris.icmb.utexas.edu/people/rong/dissertation

REFERENCES

Anachkova, B., V. Djeliova, and G. Russev. 2005. Nuclear matrix support of DNA replication. *J Cell Biochem* **96:** 951-961.

Bhat, V.B., M.H. Choi, J.S. Wishnok, and S.R. Tannenbaum. 2005. Comparative plasma proteome analysis of lymphoma-bearing SJL mice. *J Proteome Res* **4:** 1814-1825.

Gorski, S. and T. Misteli. 2005. Systems biology in the cell nucleus. *J Cell Sci* **118:** 4083-4092.

Janeway, C.A.J., P. Travers, M. Walport, and M.J. Shlomchik. 2005. *Immunobiology*. Garlan Science.

Krude, T., C. Musahl, R.A. Laskey, and R. Knippers. 1996. Human replication proteins hCdc21, hCdc46 and P1Mcm3 bind chromatin uniformly before S-phase and are displaced locally during DNA replication. *J Cell Sci* **109 (Pt 2):** 309-318.

Lea, N.C., S.J. Orr, K. Stoeber, G.H. Williams, E.W. Lam, M.A. Ibrahim, G.J. Mufti, and N.S. Thomas. 2003. Commitment point during G0-->G1 that controls entry into the cell cycle. *Mol Cell Biol* **23:** 2351-2361.

Mitulovic, G., C. Stingl, M. Smoluch, R. Swart, J.P. Chervet, I. Steinmacher, C. Gerner, and K. Mechtler. 2004. Automated, on-line two-dimensional nano liquid chromatography tandem mass spectrometry for rapid analysis of complex protein digests. *Proteomics* **4:** 2545-2557.

Nesvizhskii, A.I., A. Keller, E. Kolker, and R. Aebersold. 2003. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75:** 4646-4658.

Nickerson, J. 2001. Experimental observations of a nuclear matrix. *J Cell Sci* **114:** 463-474.

Simpson, J.C., R. Wellenreuther, A. Poustka, R. Pepperkok, and S. Wiemann. 2000. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep* **1:** 287-292.

Stein, G.S., J.B. Lian, A.J. van Wijnen, J.L. Stein, A. Javed, M. Montecino, S.K. Zaidi, D. Young, J.Y. Choi, S. Gutierrez, and S. Pockwinse. 2004. Nuclear microenvironments support assembly and organization of the transcriptional regulatory machinery for cell proliferation and differentiation. *J Cell Biochem* **91:** 287-302.

Wang, R., J.T. Prince, and E.M. Marcotte. 2005. Mass spectrometry of the M. smegmatis proteome: protein expression levels correlate with function, operons, and codon bias. *Genome Res* **15:** 1118-1126.

Washburn, M.P., D. Wolters, and J.R. Yates, 3rd. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19:** 242-247.

Yi, E.C., M. Marelli, H. Lee, S.O. Purvine, R. Aebersold, J.D. Aitchison, and D.R. Goodlett. 2002. Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis* **23:** 3205-3216.

Chapter 6. Conclusions

Proteomics is the study of all proteins expressed by the genome. Shotgun proteomics mass spectrometry has ushered in comprehensive analyses of the proteome (Ong and Mann 2005). In this work, we described application of shotgun proteomics to several biological problems, primarily the mapping of the *Mycobacterium smegmatis* proteome. Using shotgun experiments coupled with the incomplete *M. smegmatis* genome sequence, approximately 2,550 distinct proteins were identified with ~ 5% false positive identification rate. Our study therefore provides direct experimental annotation for the *M. smegmatis* genome.

Protein expression levels were calculated from the number of observations for each protein, and the number of identified peptides for each protein. In the first case, we measured differential expression of complete operons, and compared proteomes in the exponential and the stationary phases. Expression levels are correlated with proteins' codon biases and mRNA expression levels, as measured by comparison with codon adaptation indices, principal component analysis of codon frequencies, and DNA microarray data. We also compared the differential protein expression of the *M. smegmatis* in response to the three anti-tuberculosis drugs isoniazid (Sinha et al.), ethambutol (EMB), and 5-chloro-pyrazinamide (5-Cl-PZA), and elucidated the drugs' systematic effects on mycobacterial cells. In conclusion, the protein profiling of *M. smegmatis* provides us a step toward deeper understanding the mycobacteria growth and guidance for deciphering the biology of *Mycobacterium tuberculosis.* Quantitative analysis of proteome of *M. smegmatis* gives a platform for discovering the mechanism of

drug responses and a tractable system for anti-mycobacterial drug development. The shotgun technology has been a core tool for proteomics study. Although we identified 100 to 1000 proteins in each experiment, far more spectra have not been identified, which requires the improvement of the data analysis tools (Prince et al. 2004). The application of statistics in qualitative and quantitative proteomics will pave the way for the functional studies (Ong and Mann 2005).

Computational methods are helpful in predicting protein-protein interactions and assigning functions to proteins in *M. tuberculosis*. The methods of phylogenetic profiles, domain fusions, gene neighbors, and operon predictions, combined with BLASTP, discover protein functional links and characterize specific functions for the metabolic pathway and parallel pathways. However, computational methods depend on the genome sequences and knowledge for the protein functions, error is unavoidable as there are incomplete sequences and misannotated proteins. Therefore, molecular biology experiments are still required for the verification of the predicted protein functions.

The shotgun technology has also been applied in the proteomic analysis in mammalian cells. The matrix-associated proteins have been identified from the human preliminary T cells during cell activation. These proteins are involved in DNA replication, RNA transcription, splicing, and so on. Nuclear proteins have been identified from the mouse T lymphoma. Instead of analyzing entire proteomes, the analyzation of subcellular fractions of particular biological interest is a more efficient and practical strategy, as the sample complexity is reduced while the information obtained is enriched in relevance.

REFERENCES

Ong, S.E. and M. Mann. 2005. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* **1:** 252-262.

Prince, J.T., M.W. Carlson, R. Wang, P. Lu, and E.M. Marcotte. 2004. The need for a public proteomics repository. *Nat Biotechnol* **22:** 471-472.

Sinha, S., K. Kosalai, S. Arora, A. Namane, P. Sharma, A.N. Gaikwad, P. Brodin, and S.T. Cole. 2005. Immunogenic membrane-associated proteins of Mycobacterium tuberculosis revealed by proteomics. *Microbiology* **151:** 2411-2419.

Bibliography

Aebersold, R. and M. Mann. 2003. Mass spectrometry-based proteomics. Nature 422: 198-207.

Allen, T., P. Shen, L. Samsel, R. Liu, L. Lindahl, and J.M. Zengel. 1999. Phylogenetic analysis of L4-mediated autogenous control of the S10 ribosomal protein operon. J Bacteriol 181: 6124-6132.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. J Mol Biol 215: 403-410.

Anachkova, B., V. Djeliova, and G. Russev. 2005. Nuclear matrix support of DNA replication. J Cell Biochem 96: 951-961.

Arthur, J.W. and M.R. Wilkins. 2003. Using proteomics to mine genome sequences. Journal of proteome research 3: 393-402.

Banerjee, A., E. Dubnau, A. Quemard, V. Balasubramanian, K.S. Um, T. Wilson, D. Collins, G. de Lisle, and W.R. Jacobs, Jr. 1994. inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis. Science 263: 227-230.

Bennetzen, J.L. and B.D. Hall. 1982. Codon selection in yeast. J Biol Chem 257: 3026-3031.

Betts, J.C., P.T. Lukey, L.C. Robb, R.A. McAdam, and K. Duncan. 2002. Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling. Mol Microbiol 43: 717-731.

Bhat, V.B., M.H. Choi, J.S. Wishnok, and S.R. Tannenbaum. 2005. Comparative plasma proteome analysis of lymphoma-bearing SJL mice. J Proteome Res 4: 1814-1825.

Boshoff, H.I., T.G. Myers, B.R. Copp, M.R. McNeil, M.A. Wilson, and C.E. Barry, 3rd. 2004. The transcriptional responses of Mycobacterium tuberculosis to inhibitors of metabolism: novel insights into drug mechanisms of action. J Biol Chem 279: 40174-40184.

Boshoff, H.I., V. Mizrahi, and C.E. Barry, 3rd. 2002. Effects of pyrazinamide on fatty acid synthesis by whole mycobacterial cells and purified fatty acid synthase I. J Bacteriol 184: 2167-2172.

Brosch, R., A.S. Pym, S.V. Gordon, and S.T. Cole. 2001. The evolution of mycobacterial pathogenicity: clues from comparative genomics. Trends Microbiol 9: 452-458.

Bugrim, A., T. Nikolskaya, and Y. Nikolsky. 2004. Early prediction of drug metabolism and toxicity: systems biology approach and modeling. Drug Discov Today 9: 127-135.

Cabusora, L., E. Sutton, A. Fulmer, and C.V. Forst. 2005. Differential network expression during drug and stress response. Bioinformatics 21: 2898-2905.

Carpene, C., S. Bour, V. Visentin, F. Pellati, S. Benvenuti, M.C. Iglesias-Osma, M.J. Garcia-Barrado, and P. Valet. 2005. Amine oxidase substrates for impaired glucose tolerance correction. J Physiol Biochem 61: 405-419.

Chacon, O., Z. Feng, N.B. Harris, N.E. Caceres, L.G. Adams, and R.G. Barletta. 2002. Mycobacterium smegmatis D-Alanine Racemase Mutants Are Not Dependent on D-Alanine for Growth. Antimicrob Agents Chemother 46: 47-54.

Colangeli, R., D. Helb, S. Sridharan, J. Sun, M. Varma-Basil, M.H. Hazbon, R. Harbacheuski, N.J. Megjugorac, W.R. Jacobs, Jr., A. Holzenburg, J.C. Sacchettini, and D. Alland. 2005. The Mycobacterium tuberculosis iniA gene is essential for activity of an efflux pump that confers drug tolerance to both isoniazid and ethambutol. Mol Microbiol 55: 1829-1840.

Cole, S.T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry, 3rd, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M.A. Quail, M.A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J.E. Sulston, K. Taylor, S. Whitehead, and B.G. Barrell. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393: 537-544.

Corbett, E.L., C.J. Watt, N. Walker, D. Maher, B.G. Williams, M.C. Raviglione, and C. Dye. 2003. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. Arch Intern Med 163: 1009-1021.

Covert, M.W., E.M. Knight, J.L. Reed, M.J. Herrgard, and B.O. Palsson. 2004. Integrating high-throughput and computational data elucidates bacterial networks. Nature 429: 92-96.

Dandekar, T., B. Snel, M. Huynen, and P. Bork. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23: 324-328.

Date, S.V. and E.M. Marcotte. 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat Biotechnol 21: 1055-1062.

David, H.L., A. Laszlo, and N. Rastogi. 1989. Mode of action of antimycobacterial drugs. Acta Leprol 7 Suppl 1: 189-194.

de Hoog, C.L. and M. Mann. 2004. Proteomics. Annu Rev Genomics Hum Genet 5: 267-293.

Delcher, A.L., D. Harmon, S. Kasif, O. White, and S.L. Salzberg. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27: 4636-4641.

Deng, L., K. Mikusova, K.G. Robuck, M. Scherman, P.J. Brennan, and M.R. McNeil. 1995. Recognition of multiple effects of ethambutol on metabolism of mycobacterial cell envelope. Antimicrob Agents Chemother 39: 694-701.

Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863-14868.

Eisenberg, D., E.M. Marcotte, I. Xenarios, and T.O. Yeates. 2000. Protein function in the post-genomic era. Nature 405: 823-826.

Eng, J.K., A.L. McCormack, and J.R.3rd. Yates. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5: 976-989.

Enright, A.J., I. Iliopoulos, N.C. Kyrpides, and C.A. Ouzounis. 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86-90.

Flory, M.R., T.J. Griffin, D. Martin, and R. Aebersold. 2002. Advances in quantitative proteomics using stable isotope tags. Trends Biotechnol 20: S23-29.

Futcher, B., G.I. Latter, P. Monardo, C.S. McLaughlin, and J.I. Garrels. 1999. A sampling of the yeast proteome. Mol Cell Biol 19: 7357-7368.

Gerber, S.A., J. Rush, O. Stemman, M.W. Kirschner, and S.P. Gygi. 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci U S A 100: 6940-6945.

Gorski, S. and T. Misteli. 2005. Systems biology in the cell nucleus. J Cell Sci 118: 4083-4092.

Gygi, S.P., B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, and R. Aebersold. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17: 994-999.

Gygi, S.P., G.L. Corthals, Y. Zhang, Y. Rochon, and R. Aebersold. 2000. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc Natl Acad Sci U S A 97: 9390-9395.

Harth, G., P.C. Zamecnik, J.Y. Tang, D. Tabatadze, and M.A. Horwitz. 2000. Treatment of Mycobacterium tuberculosis with antisense oligonucleotides to glutamine synthetase mRNA inhibits glutamine synthetase activity, formation of the poly-L-glutamate/glutamine cell wall structure, and bacterial replication. Proc Natl Acad Sci U S A 97: 418-423.

Hughes, M.A., J.C. Silva, S.J. Geromanos, and C.A. Townsend. 2006. Quantitative proteomic analysis of drug-induced changes in mycobacteria. J Proteome Res 5: 54-63.

Ishihama, A. 2000. Functional modulation of Escherichia coli RNA polymerase. Annu Rev Microbiol 54: 499-518.

Ishihama, Y., Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, and M. Mann. 2005. Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. Mol Cell Proteomics 4: 1265-1272.

Jacobs, W.R., Jr., G.V. Kalpana, J.D. Cirillo, L. Pascopella, S.B. Snapper, R.A. Udani, W. Jones, R.G. Barletta, and B.R. Bloom. 1991. Genetic systems for mycobacteria. Methods Enzymol 204: 537-555.

Jaffe, J.D., H.C. Berg, and G.M. Church. 2004. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4: 59-77.

Janeway, C.A.J., P. Travers, M. Walport, and M.J. Shlomchik. 2005. Immunobiology. Garlan Science.

Jansen, R., H.J. Bussemaker, and M. Gerstein. 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Res 31: 2242-2251.

Jolliffe, I.T. 2002. Principal component analysis. Springer.

Jungblut, P.R., E.C. Muller, J. Mattow, and S.H. Kaufmann. 2001. Proteomics reveals open reading frames in Mycobacterium tuberculosis H37Rv not predicted by genomics. Infect Immun 69: 5905-5907.

Jungblut, P.R., U.E. Schaible, H.J. Mollenkopf, U. Zimny-Arndt, B. Raupach, J. Mattow, P. Halada, S. Lamer, K. Hagens, and S.H. Kaufmann. 1999. Comparative proteome analysis of Mycobacterium tuberculosis and Mycobacterium bovis BCG strains: towards functional genomics of microbial pathogens. Mol Microbiol 33: 1103-1117.

Kanehisa, M. and S. Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30.

Keller, A., A.I. Nesvizhskii, E. Kolker, and R. Aebersold. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74: 5383-5392.

Kormanec, J., B. Sevcikova, N. Halgasova, R. Knirschova, and B. Rezuchova. 2000. Identification and transcriptional characterization of the gene encoding the stress-response sigma factor sigma(H) in streptomyces coelicolor A3(2). FEMS Microbiol Lett 189: 31-38.

Krude, T., C. Musahl, R.A. Laskey, and R. Knippers. 1996. Human replication proteins hCdc21, hCdc46 and P1Mcm3 bind chromatin uniformly before S-phase and are displaced locally during DNA replication. J Cell Sci 109 (Pt 2): 309-318.

Lea, N.C., S.J. Orr, K. Stoeber, G.H. Williams, E.W. Lam, M.A. Ibrahim, G.J. Mufti, and N.S. Thomas. 2003. Commitment point during G0-->G1 that controls entry into the cell cycle. Mol Cell Biol 23: 2351-2361.

Lee, I., S.V. Date, A.T. Adai, and E.M. Marcotte. 2004. A probabilistic functional network of yeast genes. Science 306: 1555-1558.

Link, A.J., J. Eng, D.M. Schieltz, E. Carmack, G.J. Mize, D.R. Morris, B.M. Garvik, and J.R. Yates, 3rd. 1999. Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol 17: 676-682.

Liu, H., R.G. Sadygov, and J.R. Yates, 3rd. 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76: 4193-4201.

Lu, P., C. Vogel, and E.M. Marcotte. 2005. An estimate of relative contributions of transcriptional and translational regulation by absolute protein expression profiling. submitted.

Marcotte, E.M. 2000. Computational genetics: finding protein function by nonhomology methods. Curr Opin Struct Biol 10: 359-365.

Marcotte, E.M., M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science 285: 751-753.

Marcotte, E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. 1999. A combined algorithm for genome-wide prediction of protein function. Nature 402: 83-86.

Mattow, J., U.E. Schaible, F. Schmidt, K. Hagens, F. Siejak, G. Brestrich, G. Haeselbarth, E.C. Muller, P.R. Jungblut, and S.H. Kaufmann. 2003. Comparative proteome analysis of culture supernatant proteins from virulent Mycobacterium tuberculosis H37Rv and attenuated M. bovis BCG Copenhagen. Electrophoresis 24: 3405-3420.

Mawuenyega, K.G., C.V. Forst, K.M. Dobos, J.T. Belisle, J. Chen, E.M. Bradbury, A.R. Bradbury, and X. Chen. 2005. Mycobacterium tuberculosis functional network analysis by global subcellular protein profiling. Mol Biol Cell 16: 396-404.

Mdluli, K., R.A. Slayden, Y. Zhu, S. Ramaswamy, X. Pan, D. Mead, D.D. Crane, J.M. Musser, and C.E. Barry, 3rd. 1998. Inhibition of a Mycobacterium tuberculosis beta-ketoacyl ACP synthase by isoniazid. Science 280: 1607-1610.

Mikusova, K., R.A. Slayden, G.S. Besra, and P.J. Brennan. 1995. Biogenesis of the mycobacterial cell wall and the site of action of ethambutol. Antimicrob Agents Chemother 39: 2484-2489.

Mitulovic, G., C. Stingl, M. Smoluch, R. Swart, J.P. Chervet, I. Steinmacher, C. Gerner, and K. Mechtler. 2004. Automated, on-line two-dimensional nano liquid chromatography tandem mass spectrometry for rapid analysis of complex protein digests. Proteomics 4: 2545-2557.

Mukhopadhyay, B. and E. Purwantini. 2000. Pyruvate carboxylase from Mycobacterium smegmatis: stabilization, rapid purification, molecular and biochemical characterization and regulation of the cellular level. Biochim Biophys Acta 1475: 191-206.

Nesvizhskii, A.I., A. Keller, E. Kolker, and R. Aebersold. 2003. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75: 4646-4658.

Nickerson, J. 2001. Experimental observations of a nuclear matrix. J Cell Sci 114: 463-474.

Oda, Y., K. Huang, F.R. Cross, D. Cowburn, and B.T. Chait. 1999. Accurate quantitation of protein expression and site-specific phosphorylation. Proc Natl Acad Sci U S A 96: 6591-6596.

Ong, S.E. and M. Mann. 2005. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol 1: 252-262.

Ong, S.E., B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, and M. Mann. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 1: 376-386.

Overbeek, R., M. Fonstein, M. D'Souza, G.D. Pusch, and N. Maltsev. 1999. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96: 2896-2901.

Pandey, A. and M. Mann. 2000. Proteomics to study genes and genomes. Nature 405: 837-846.

Parish, T. and N.G. Stoker. 1998. Mycobacterial Protocols. Humana Press. Humana press.

Parish, T. and N.G. Stoker. 2001. Mycobacterium tuberculosis Protocols. Humana press.

Pellegrini, M., E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285-4288.

Peng, J., J.E. Elias, C.C. Thoreen, L.J. Licklider, and S.P. Gygi. 2003. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res 2: 43-50.

Perkins, D.N., D.J. Pappin, D.M. Creasy, and J.S. Cottrell. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20: 3551-3567.

Phetsuksiri B, B.A., Cooper AM, Minnikin DE, Douglas JD, Besra GS, Brennan PJ. 1999. Antimycobacterial activities of isoxyl and new derivatives through the inhibition of mycolic acid synthesis. Antimicrob Agents Chemother. 43: 1042-1051.

Primm, T.P., S.J. Andersen, V. Mizrahi, D. Avarbock, H. Rubin, and C.E. Barry, 3rd. 2000. The stringent response of Mycobacterium tuberculosis is required for long-term survival. J Bacteriol 182: 4889-4898.

Prince, J.T., M.W. Carlson, R. Wang, P. Lu, and E.M. Marcotte. 2004. The need for a public proteomics repository. Nat Biotechnol 22: 471-472.

Radmacher, E., K.C. Stansen, G.S. Besra, L.J. Alderwick, W.N. Maughan, G. Hollweg, H. Sahm, V.F. Wendisch, and L. Eggeling. 2005. Ethambutol, a cell wall inhibitor of Mycobacterium tuberculosis, elicits L-glutamate efflux of Corynebacterium glutamicum. Microbiology 151: 1359-1368.

Raman, K., P. Rajagopalan, and N. Chandra. 2005. Flux balance analysis of mycolic Acid pathway: targets for anti-tubercular drugs. PLoS Comput Biol 1: e46.

Rapaport, E., A. Levina, V. Metelev, and P.C. Zamecnik. 1996. Antimycobacterial activities of antisense oligodeoxynucleotide phosphorothioates in drug-resistant strains. Proc Natl Acad Sci U S A 93: 709-713.

Ratledge, C.a.D., J. 1999. Mycobacteria molecular biology and virulence. Blackwell science.

Raychaudhuri, S., M. Basu, and N.C. Mandal. 1998. Glutamate and cyclic AMP regulate the expression of galactokinase in Mycobacterium smegmatis. Microbiology 144 (Pt 8): 2131-2140.

Rison, S.C., S.A. Teichmann, and J.M. Thornton. 2002. Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in Escherichia coli. J Mol Biol 318: 911-932.

Saito, K., S. Fujigaki, M.P. Heyes, K. Shibata, M. Takemura, H. Fujii, H. Wada, A. Noma, and M. Seishima. 2000. Mechanism of increases in L-kynurenine and quinolinic acid in renal insufficiency. Am J Physiol Renal Physiol 279: F565-572.

Salgado, H., G. Moreno-Hagelsieb, T.F. Smith, and J. Collado-Vides. 2000. Operons in Escherichia coli: genomic analyses and predictions. Proc Natl Acad Sci U S A 97: 6652-6657.

Schmidt, F., S. Donahoe, K. Hagens, J. Mattow, U.E. Schaible, S.H. Kaufmann, R. Aebersold, and P.R. Jungblut. 2004. Complementary analysis of the Mycobacterium tuberculosis proteome by two-dimensional electrophoresis and isotope-coded affinity tag technology. Mol Cell Proteomics 3: 24-42.

Scorpio, A. and Y. Zhang. 1996. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. Nat Med 2: 662-667.

Sharp, P.M. and W.H. Li. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15: 1281-1295.

Shibata, K., T. Fukuwatari, and E. Sugimoto. 2001. Effects of dietary pyrazinamide, an antituberculosis agent, on the metabolism of tryptophan to niacin and of tryptophan to serotonin in rats. Biosci Biotechnol Biochem 65: 1339-1346.

Simpson, J.C., R. Wellenreuther, A. Poustka, R. Pepperkok, and S. Wiemann. 2000. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. EMBO Rep 1: 287-292.

Sinha, S., K. Kosalai, S. Arora, A. Namane, P. Sharma, A.N. Gaikwad, P. Brodin, and S.T. Cole. 2005. Immunogenic membrane-associated proteins of Mycobacterium tuberculosis revealed by proteomics. Microbiology 151: 2411-2419.

Stein, G.S., J.B. Lian, A.J. van Wijnen, J.L. Stein, A. Javed, M. Montecino, S.K. Zaidi, D. Young, J.Y. Choi, S. Gutierrez, and S. Pockwinse. 2004. Nuclear microenvironments support assembly and organization of the transcriptional regulatory machinery for cell proliferation and differentiation. J Cell Biochem 91: 287-302.

Tabb, D.L., McDonald, W.H., and Yates III, J.R. 2002. DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. 1: 21–26.

Takayama, K. and J.O. Kilburn. 1989. Inhibition of synthesis of arabinogalactan by ethambutol in Mycobacterium smegmatis. Antimicrob Agents Chemother 33: 1493-1499.

Takayama, K., C. Wang, and G.S. Besra. 2005. Pathway to synthesis and processing of mycolic acids in Mycobacterium tuberculosis. Clin Microbiol Rev 18: 81-101.

Talaat, A.M., S.T. Howard, W.t. Hale, R. Lyons, H. Garner, and S.A. Johnston. 2002. Genomic DNA standards for gene expression profiling in Mycobacterium tuberculosis. Nucleic Acids Res 30: e104.

Tatusov, R.L., D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, and E.V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29: 22-28.

Thackray, P.D. and A. Moir. 2003. SigM, an extracytoplasmic function sigma factor of Bacillus subtilis, is activated in response to cell wall antibiotics, ethanol, heat, acid, and superoxide stress. J Bacteriol 185: 3491-3498.

Tringe, S.G., C. von Mering, A. Kobayashi, A.A. Salamov, K. Chen, H.W. Chang, M. Podar, J.M. Short, E.J. Mathur, J.C. Detter, P. Bork, P. Hugenholtz, and E.M. Rubin. 2005. Comparative metagenomics of microbial communities. Science 308: 554-557.

Vitale, G., R. Pellizzari, C. Recchi, G. Napolitani, M. Mock, and C. Montecucco. 1998. Anthrax lethal factor cleaves the N-terminus of MAPKKs and induces tyrosine/threonine phosphorylation of MAPKs in cultured macrophages. Biochem Biophys Res Commun 248: 706-711.

Wade, M.M. and Y. Zhang. 2004. Anaerobic incubation conditions enhance pyrazinamide activity against Mycobacterium tuberculosis. J Med Microbiol 53: 769-773.

Wang, R., J.T. Prince, and E.M. Marcotte. 2005. Mass spectrometry of the M. smegmatis proteome: protein expression levels correlate with function, operons, and codon bias. Genome Res 15: 1118-1126.

Washburn, M.P. and J.R. Yates, 3rd. 2000. Analysis of the microbial proteome. Curr Opin Microbiol 3: 292-297.

Washburn, M.P., D. Wolters, and J.R. Yates, 3rd. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19: 242-247.

Wilson, M., J. DeRisi, H.H. Kristensen, P. Imboden, S. Rane, P.O. Brown, and G.K. Schoolnik. 1999. Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization. Proc Natl Acad Sci U S A 96: 12833-12838.

Wu, S., S.T. Howard, D.L. Lakey, A. Kipnis, B. Samten, H. Safi, V. Gruppo, B. Wizel, H. Shams, R.J. Basaraba, I.M. Orme, and P.F. Barnes. 2004. The principal sigma factor sigA mediates enhanced growth of Mycobacterium tuberculosis in vivo. Mol Microbiol 51: 1551-1562.

Yeung, K.Y. and W.L. Ruzzo. 2001. Principal component analysis for clustering gene expression data. Bioinformatics 17: 763-774.

Yi, E.C., M. Marelli, H. Lee, S.O. Purvine, R. Aebersold, J.D. Aitchison, and D.R. Goodlett. 2002. Approaching complete peroxisome characterization by gas-phase fractionation. Electrophoresis 23: 3205-3216.

Yu, J., L. Hederstedt, and P.J. Piggot. 1995. The cytochrome bc complex (menaquinone:cytochrome c reductase) in Bacillus subtilis has a nontraditional subunit organization. J Bacteriol 177: 6751-6760.

Zahrt, T.C. and V. Deretic. 2000. An essential two-component signal transduction system in Mycobacterium tuberculosis. J Bacteriol 182: 3832-3838.

Zhang, Y. 2005. The magic bullets and tuberculosis drug targets. Annu Rev Pharmacol Toxicol 45: 529-564.

Zhang, Y. and D. Mitchison. 2003. The curious characteristics of pyrazinamide: a review. Int J Tuberc Lung Dis 7: 6-21.

Zimhony, O., J.S. Cox, J.T. Welch, C. Vilcheze, and W.R. Jacobs, Jr. 2000. Pyrazinamide inhibits the eukaryotic-like fatty acid synthetase I (FASI) of Mycobacterium tuberculosis. Nat Med 6: 1043-1047.

Vita

Rong Wang was born in Tianjin, China, on September 9, 1974, the daughter of Shulan Wang and Wenling Wang. In 1992, she completed her high-school degree from Tanggu First High School in Tianjin with the Prize in National Chemistry Competition and was accepted to Fudan University in Shanghai. She received her Bachelor's degree in Biochemistry and Master's degree from the Genetics Institute of Fudan University. In the fall semester of 2000, she came to the University of Texas at Austin as a graduate student in the Cellular and Molecular Biology program. She joined Dr. Edward Marcotte's lab to study mycobacteria proteomics. Her current scientific publications include:

Wang, R., and Marcotte, E.M. Shotgun proteomic analysis of drug treated mycobacteria. manuscript in preparation.

Wang, R., Prince, J.T. and Marcotte, E.M. Mass-spectrometry of the *M. smegmatis* proteome: protein expression levels correlate with function, operons, and codon bias. *Genome Research.* 2005, 15, 8:1118-1126.

Prince, J.T., Carlson, M.W., Wang, R., Lu P. and Marcotte, E.M. The need for a public proteomic repository. *Nat Biotechnol.* 2004, 22 (4): 471-472.

Wang, R., Yu H., Zhou W.G., Li, C.B., and Zhao, S.Y. Study on Expression of a Chimeric Molecule B7-TNFR in Human Tumor Cell Line. *Journal of FUDAN University, Natural Science* 2000, 3, 39: 259-263.

Zhou, W.G., Zhao, X.Y., Wang, R., Gan, B.Y., Li, C.B., and Zhao, S.Y. The Infection of Wolbachia in *Drosophila auraria* and *Drosophila simulans*. *Journal of FUDAN University, Natural Science* 1999, 4, 38: 251-256.

Permanent address: Xinhua Rd. Xinhuadongli, 3-2-301, Tanggu, Tianjin, China. 300450

This dissertation was typed by the Author.